



**HAL**  
open science

# Modélisation et Algorithmique de graphes pour la construction de structures moléculaires.

Marie Bricage

► **To cite this version:**

Marie Bricage. Modélisation et Algorithmique de graphes pour la construction de structures moléculaires.. Géométrie algorithmique [cs.CG]. Université Paris Saclay (COmUE), 2018. Français. NNT : 2018SACLV031 . tel-01955838

**HAL Id: tel-01955838**

**<https://theses.hal.science/tel-01955838v1>**

Submitted on 14 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Modélisation et Algorithmique de graphes pour la construction de structures moléculaires

Thèse de doctorat de l'Université Paris-Saclay  
préparée à l'Université de Versailles Saint-Quentin

Ecole doctorale n°580 Sciences et technologies de l'information et de la  
communication (STIC)  
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Versailles, le 5 Juin 2018, par

**MME MARIE BRICAGE**

Composition du Jury :

Mme Alessandra Carbone Professeur, Université Pierre et Marie Curie	Président
M. Matthieu Latapy Directeur de recherche, Sorbonne Université	Rapporteur
M. Stéphane Vialette Directeur de recherche, Université Paris-Est Marne-la-Vallée	Rapporteur
M. Dominique Barth Professeur, Université de Versailles Saint-Quentin	Directeur de thèse
M. Olivier David Maître de conférences, Université de Versailles Saint-Quentin	Examineur
Mme Sandrine Vial Maître de conférence, Université de Versailles Saint-Quentin	Examineur



**Mots clés :** Algorithmique, molécules, graphes, modélisation

**Résumé :** Dans cette thèse, nous présentons une approche algorithmique permettant la génération de guides de construction de cages moléculaires organiques. Il s'agit d'architectures 3D semi-moléculaires possédant un espace interne défini capable de piéger une molécule cible appelée substrat. De nombreuses œuvres proposent de générer des cages organiques moléculaires obtenues à partir de structures symétriques, qui ont une bonne complexité, mais elles ne sont pas spécifiques car elles ne prennent pas en compte des cibles précises. L'approche proposée permet de générer des guides de construction de cages moléculaires organiques spécifiques à un substrat donné.

Afin de garantir la spécificité de la cage moléculaire pour le substrat cible, une structure intermédiaire, qui est une expansion de l'enveloppe du substrat cible, est utilisée. Cette structure définit la forme de l'espace dans lequel est piégé le substrat.

Bien que cette structure ne soit pas assimilable à une molécule, elle est construite dans une logique moléculaire. Chaque sommet qui la compose est situé dans une direction dans laquelle l'un des atomes du substrat peut créer des interactions avec des atomes d'autres molécules. Pour ce faire, nous utilisons les propriétés des éléments chimiques qui composent le substrat, ainsi que la méthode de VSEPR. Cette méthode permet de prédire la géométrie des atomes en se basant sur la théorie de la répulsion des électrons. À partir de leur géométrie, pour chaque atome du substrat, on peut retrouver les directions des doublets non-liants d'un atome. Les directions de ces doublets non-liants sont les directions dans lesquelles un atome peut créer des interactions (liaisons) avec d'autres atomes.

Des petits ensembles d'atomes, appelés motifs moléculaires liants, sont ensuite intégrés à cette structure intermédiaire. Ces motifs moléculaires sont les ensembles d'atomes nécessaires aux cages moléculaires pour leur permettre d'interagir avec le substrat afin de le capturer.

**Title :** Modelling and graph algorithms for building molecular structures.

**Keywords :** Algorithmic, Molecules, Graphs, Modelling

**Summary :** In this thesis, we present an algorithmic approach allowing the generation of construction guides of organic molecular cages. These semi-molecular architectures have a defined internal space capable of trapping a target molecule called substrate. Many works propose to generate molecular organic cages obtained from symmetrical structures, which have a good complexity, but they are not specific because they do not take into account precise targets. The proposed approach makes it possible to generate guides for the construction of organic molecular cages specific to a given substrate.

In order to ensure the specificity of the molecular cage for the target substrate, an intermediate structure, which is an expansion of the envelope of the target substrate, is used. This structure defines the shape of the space in which the substrate is trapped.

Although, this structure is not comparable to a molecule, but it is built in a molecular logic. Each vertex which composes it, is located in a direction in which one of the atoms of the substrate can create interactions with atoms of other molecules. To do this, we use the properties of the chemical elements that make up the substrate, as well as the VSEPR method. This method predicts the geometry of atoms based on the electron repulsion theory. From their geometry, for each atom of the substrate, we can find the directions of the lone pairs of an atom. The directions of these lone pairs are the directions in which an atom can create interactions (bonds) with other atoms.

Small sets of atoms, called molecular binding patterns, are then integrated into this intermediate structure. These molecular patterns are the sets of atoms needed by molecular cages to allow them to interact with the substrate to capture it.

# Table des matières

Table des matières	v
Table des figures	vii
Table des tableaux	ix
Introduction	1
<b>1 Contexte</b>	<b>5</b>
1 Contexte de la thèse . . . . .	6
1.1 Rappel de chimie . . . . .	6
1.2 Chimie supra-moléculaire . . . . .	7
1.3 Objectif : Cages moléculaires . . . . .	9
2 Apparition et utilisation des cages moléculaires . . . . .	10
3 Approche proposée . . . . .	12
3.1 Contraintes des graphes moléculaires . . . . .	12
3.2 Motifs moléculaires . . . . .	15
3.3 Idée générale . . . . .	15
4 Conclusion . . . . .	17
<b>2 Construction de l'enveloppe</b>	<b>19</b>
1 Données et Modélisation . . . . .	20
1.1 Données d'entrée . . . . .	20
1.2 Enveloppe . . . . .	20
2 Topologie des sommets de $G$ . . . . .	21
2.1 Topologie des sommets ayant deux voisins ou plus . . . . .	21
2.2 Topologie des sommets possédant un voisin unique . . . . .	28
2.3 Algorithme général pour la recherche de la topologie d'un sommet . . . . .	29
3 Expansion du substrat . . . . .	30
3.1 Normale d'un sommet . . . . .	31
3.2 Extension des sommets à géométrie tétraédrique . . . . .	32
3.3 Extension des sommets à géométrie triangulaire . . . . .	34
3.4 Extension des sommets à géométrie linéaire . . . . .	36
3.5 Récapitulatif . . . . .	37
4 Construction des arêtes de l'enveloppe . . . . .	38
4.1 Propriété attendue . . . . .	38
4.2 Alpha shape . . . . .	39
5 Conclusion . . . . .	41
<b>3 Intégration des motifs liants</b>	<b>43</b>
1 Intégration d'un motif moléculaire . . . . .	44
1.1 Étapes d'intégration d'un motif . . . . .	44

---

1.2	Positionnement d'un motif et recouvrement . . . . .	48
1.3	Bordure et rattachement . . . . .	51
2	Construction des liaisons aromatiques . . . . .	51
3	Construction des Liaisons Hydrogènes . . . . .	54
3.1	Donneur / Accepteur . . . . .	55
3.2	Graphes des dépendances . . . . .	56
3.3	Intégration des liaisons hydrogènes . . . . .	59
4	Conclusion . . . . .	60
<b>4</b>	<b>Etude de cas</b>	<b>63</b>
1	Description de l'application . . . . .	64
2	Étude de l'approche sur la molécule l'adénosine . . . . .	64
2.1	Construction de l'enveloppe . . . . .	65
2.2	Insertion des motifs liants . . . . .	69
3	Étude d'approche sur la molécule de saccharose . . . . .	73
3.1	Construction de l'enveloppe . . . . .	73
3.2	Insertion des liaisons hydrogènes . . . . .	73
4	Étude de l'approche sur les molécules d'acétanilide et D-tyrosine . . . .	76
4.1	Exemples de la molécule d'acétanilide . . . . .	77
4.2	Exemples de la molécule de D-tyrosine . . . . .	80
5	Conclusion . . . . .	83
	<b>Conclusion</b>	<b>87</b>
<b>5</b>	<b>Bibliographie</b>	<b>91</b>
<b>A</b>	<b>Rappel de chimie : Règle du duet et règle de l'octet</b>	<b>95</b>
<b>B</b>	<b>Étude de l'influence du paramètre <math>\alpha</math> de la méthode alphashape</b>	<b>97</b>

# Table des figures

1.1	Exemples atomes et de leurs couches électroniques. . . . .	6
1.2	Schématisation de liaisons covalentes et répartition des électrons autour des atomes liés. . . . .	7
1.3	Schématisation d'une liaison hydrogène entre deux molécules d'eau. . .	8
1.4	Schématisation d'une molécule de benzène. . . . .	9
1.5	Exemple de compatibilité entre une cage et un substrat. . . . .	10
1.6	Exemples de spécificité entre une cage et un substrat. . . . .	10
1.7	Récapitulatif des phases de la génération de guide par les cages moléculaires d'un substrat donné. . . . .	17
2.1	Exemple de réduction d'un graphe. . . . .	23
2.2	Chemins trouvés pour un sommet du graphe $R$ appartenant à un cycle. . . . .	25
2.3	Chemins trouvés pour un sommet du graphe $R$ n'appartenant pas à un cycle. . . . .	25
2.4	Exemples de sous-graphe de $G$ . . . . .	26
2.5	Sommets à géométrie tétraédrique . . . . .	27
2.6	Sommets à géométrie triangulaire . . . . .	27
2.7	Sommet à géométrie linéaire de topologie $(2, 0)$ . . . . .	27
2.8	Cycle régulier à 5 sommets. . . . .	28
2.9	Topologie des sommets ne possédant qu'un seul voisin . . . . .	29
2.10	Position du vecteur $normal_v$ en fonction des propriétés des sommets. . . . .	32
2.11	Directions libres des sommets de topologie $(4, 0)$ et $(3, 1)$ . . . . .	33
2.12	Directions libres d'un sommet de topologie $(2, 2)$ . . . . .	34
2.13	Directions libres d'un sommet de topologie $(1, 3)$ . . . . .	35
2.14	Directions libres des sommets à géométrie triangulaire. . . . .	36
2.15	Directions libres des sommets à géométrie linéaire. . . . .	37
2.16	Exemple d'enveloppe d'un nuage de points en 2D. . . . .	39
2.17	Exemple d'une étape de la triangulation de Delaunay. . . . .	40
2.18	Représentation de la méthode <i>Alpha-shape</i> . . . . .	41
3.1	Exemple de positionnement d'un motif dans une enveloppe. . . . .	45
3.2	Exemple des étapes 1 et 2 de l'intégration d'un motif. . . . .	45
3.3	Exemple des étapes 3 et 4 de l'intégration d'un motif. . . . .	46
3.4	Exemple de deux positionnements d'un même motif dans une enveloppe. . . . .	47
3.5	Exemple des étapes 1 et 2 de l'intégration d'un motif avec recouvrement. . . . .	47
3.6	Exemple des étapes 3 et 4 de l'intégration d'un motif avec recouvrement. . . . .	48
3.7	Exemples de positionnement d'un motif en fonction de son sommet <i>central</i> . . . . .	49
3.8	Exemple de positionnement de deux motifs différents en fonction de la direction libre d'un sommet d'intégration. . . . .	49

3.9	Exemples de positionnement d'un même motif en fonction du vecteur <i>normal</i> d'un sommet d'intégration. . . . .	50
3.10	Exemple d'extension d'un composé aromatique . . . . .	52
3.11	Motifs aromatiques complets. . . . .	53
3.12	intégration d'un cycle complet. . . . .	54
3.13	Graphes de dépendance. . . . .	58
4.1	Molécule d'adénosine . . . . .	65
4.2	Extensions des sommets du cycle de la partie ribose de la molécule d'adénosine. . . . .	67
4.3	Extensions du cycle composé de six sommets de la molécule d'adénosine. . . . .	68
4.4	Extensions des sommets du cycle composé de cinq sommets de la partie adénine de la molécule d'adénosine. . . . .	68
4.5	Construction des arêtes de l'enveloppe avec $\alpha = 4$ . . . . .	69
4.6	Reconstruction des cycles aromatiques de l'enveloppe. Les sommets fixés des motifs sont en orange et les autres sommets des motifs sont en rouge. . . . .	70
4.7	Molécule d'adénosine et son graphe de dépendance . . . . .	71
4.8	Graphes de dépendance de la molécule d'adénosine et celui de son enveloppe . . . . .	72
4.9	Exemple d'une enveloppe après insertion de motifs hydrogènes. Les sommets fixés sont en bleu et les autres sommets des motifs sont en rouge. . . . .	73
4.10	Molécule de saccharose . . . . .	74
4.11	Enveloppe de la molécule de saccharose. . . . .	75
4.12	Graphe de dépendance de la molécule de saccharose et les extensions conservées dans l'enveloppe de ces sommets. . . . .	76
4.13	Graphe de dépendance de l'enveloppe de saccharose. . . . .	76
4.14	Exemple d'une enveloppe après insertion de motifs hydrogènes. . . . .	77
4.15	Molécule d'acétanilide . . . . .	78
4.16	Enveloppe de l'acétanilide . . . . .	79
4.17	Reconstruction des cycles aromatiques de l'enveloppe. . . . .	79
4.18	Extensions des sommets du graphe de dépendance de la molécule d'acétanilide et graphe de dépendance de son enveloppe . . . . .	80
4.19	Exemple d'une enveloppe après insertion de motifs hydrogènes . . . . .	80
4.20	Molécule de D-tyrosine . . . . .	81
4.21	Enveloppe de la molécule de D-tyrosine . . . . .	82
4.22	Reconstruction des cycles aromatiques de l'enveloppe. . . . .	83
4.23	Molécule de D-tyrosine et son graphe de dépendance . . . . .	83
4.24	Molécule de D-tyrosine et graphe de dépendance de son enveloppe . . . . .	84
4.25	Exemple d'une enveloppe après insertion de motifs hydrogènes . . . . .	84
B.1	Enveloppe de la molécule d'adénosine obtenue avec $\alpha = 1$ . . . . .	97
B.2	Enveloppe de la molécule d'adénosine obtenue avec $\alpha = 2$ . . . . .	98

# Table des tableaux

1.1	Rayons de covalence . . . . .	13
1.2	Représentation de VSEPR en fonction du nombre de doublets des atomes.	14
3.1	Modifs moléculaires des liaisons hydrogènes . . . . .	55
3.2	Exemple d'application de l'algorithme 8 . . . . .	61
4.1	Topologie des atomes de la molécule d'adénosine. . . . .	66
4.2	Topologie des atomes de la molécule de saccharose. . . . .	75
4.3	Topologie des atomes de la molécules d'acétanilide. . . . .	77
4.4	Topologie des atomes de la molécules de D-tyrosine. . . . .	82



# Introduction

Les cages moléculaires constituent un enjeu majeur dans de nombreux domaines. Allant de la dépollution à la vectorisation de médicaments, en passant par le transport ou le stockage de produits dangereux, leur utilisation est très diversifiée. Les cages moléculaires sont des molécules capables de reconnaître et de capturer d'autres molécules. C'est un système hôte/invité. Elles peuvent également avoir des propriétés, c'est par exemple le cas des enzymes qui ont des propriétés catalytiques. La conception de cages moléculaires est un enjeu important dans ces domaines. Qu'elles soient construites pour stabiliser ou capturer d'autres molécules, ou encore comme enzymes artificielles pour remplacer ou copier des enzymes présentes dans la nature, leur conception dépend des utilisations pour lesquelles elles sont créées.

Par exemple certains produits chimiques sont dangereux, car ils sont instables au contact de l'oxygène de l'air, de l'eau ou d'autres substances présentes dans l'atmosphère ; dans ce cas, ils peuvent exploser ou former des gaz inflammables ou toxiques. Il est donc difficile de les stocker ou de les déplacer. En fabriquant des cages moléculaires capables de capturer et de stabiliser les molécules de ces produits chimiques, on limite les risques d'explosion ou d'inflammation. C'est le cas par exemple du phosphore blanc qui est essentiellement utilisé par les militaires comme agent incendiaire, agent de protection par écran de fumée et comme un composant d'arme anti-personnel capable de provoquer de graves brûlures. Prasenjit Mal, du Département de chimie de l'Université de Cambridge (Royaume-Uni), et ses collègues ont fabriqué une cage capable de le piéger et de le stabiliser, diminuant ainsi les risques lors de son stockage ou de son déplacement.

Un autre exemple d'utilisation de cages moléculaires est l'encapsulation de substances thérapeutiques actives. Il peut arriver que certaines molécules présentes dans des crèmes puissent devenir irritantes pour la peau si elles réagissent au préalable avec l'extérieur. En encapsulant ses molécules avec des cages moléculaires, on les stabilise afin qu'elles n'aient pas d'interactions extérieures avant utilisation.

Bien que les études sur les cages moléculaires aient débuté dès les années 1920, l'engouement autour de ces structures s'est particulièrement développé après le prix Nobel de chimie de 1987 ([Cram, 1987](#); [Lehn, 1987](#); [Pedersen, 1987](#)). Depuis l'intérêt qu'on leur porte n'a été que grandissant, en particulier grâce à leurs nombreuses applications.

Dans cette thèse, nous nous intéressons aux problèmes de conception de cages moléculaires. Plus précisément, nous proposons une contribution à la construction de cages moléculaires à partir d'un substrat donné. Afin qu'une molécule soit considérée comme une cage moléculaire pour un substrat, elle doit être capable de le capturer, c'est-à-dire d'interagir avec lui en créant des liaisons faibles entre eux telles que des

liaisons hydrogènes. Plus la molécule cage a une forme géométrique proche de celle du substrat, plus elle est spécifique à celui-ci. Par exemple, si on utilise une grosse cage moléculaire avec une forme très générique telle qu'une boule, cette cage pourra capturer toutes sortes de substrats du moment que la forme de celui-ci peut être inclus dans la forme de la cage. Cette cage ne sera donc pas considérée comme spécifique aux substrats puisqu'elle pourra en capturer un grand nombre. À l'inverse si la forme de la cage est très proche de celle d'un substrat donné, le nombre de substrats dont la forme pourra être inclus dans la forme de la cage sera sensiblement réduit rendant ainsi la cage plus spécifique au substrat donné. De même plus il y aura de liaisons faibles qui pourront s'établir entre une cage moléculaire et un substrat plus celui-ci sera spécifique.

Ici, nous avons travaillé sur une méthode permettant, à partir d'un substrat donné, de générer des guides de construction de cages moléculaires afin qu'elles soient les plus spécifiques possibles. Pour ce faire nous proposons de passer par une structure intermédiaire construite à partir du substrat qui sert par la suite de fondation pour générer les cages moléculaires. Cette structure que nous appelons *enveloppe* peut être considérée comme une « seconde peau » du substrat puisqu'elle l'englobe et a une forme complémentaire à la surface extérieure du substrat. Cependant un espace équivalent à la distance d'une liaison faible est présent entre l'enveloppe et le substrat. L'enveloppe définit donc une « zone interdite » pour les sommets des cages moléculaires générées pour conserver une cohérence chimique. De plus les sommets qui constituent l'enveloppe ne sont pas placés au hasard. En effet chacun d'entre eux est placé dans une des directions où un atome du substrat peut encore créer des interactions avec des atomes d'autres molécules. Les sommets de l'enveloppe donnent ainsi des indications sur les emplacements où pourront être positionnés des sommets des cages moléculaires pour permettre la création de liaisons faibles entre les deux molécules. Les guides des cages moléculaires seront construites par assemblages de modules moléculaires actifs, c'est-à-dire en assemblant des morceaux de molécules les uns avec les autres. L'idée est de positionner les modules moléculaires des cages qui interagiront avec ceux du substrat (motifs liants).

La première partie du document décrira les connaissances de base sur les molécules et les cages moléculaires. Elle établira également un profil des études déjà réalisées sur le sujet des cages moléculaires et décrira les différences avec l'approche que nous proposons ici.

La seconde partie du document se concentrera sur les différentes étapes de la construction de l'enveloppe du substrat. Elle décrira pour chaque atome comment nous déterminons les directions dans lesquelles ils peuvent encore avoir des interactions avec des atomes d'autres molécules.

La troisième partie du document montrera comment intégrer dans l'enveloppe des petits ensembles d'atomes pouvant interagir avec les atomes du substrat. Ces petits ensembles d'atomes sont appelés motifs moléculaires liants. Chaque motif liant intégré dans l'enveloppe sera un morceau de cage moléculaire. Ils permettront la création de liaisons entre le substrat et les cages moléculaires qui les contiendront.

Enfin la dernière partie du document présentera les résultats obtenus jusqu'à maintenant sur différents substrats. Les premiers résultats proposés sont ceux de la molécule d'adénosine qui met en avant toutes les caractéristiques du modèle. Le second

exemple est celui de la molécule de sucrose qui a pour particularité de pouvoir créer de nombreuses liaisons hydrogènes. Enfin nous comparerons les résultats de la molécule d'acétanilide et de la molécule D-tyrosine qui ont des structures assez proches mais avec des géométries différentes.



# Chapitre 1

## Contexte

---

1	Contexte de la thèse . . . . .	6
2	Apparition et utilisation des cages moléculaires . . . . .	10
3	Approche proposée . . . . .	12
4	Conclusion . . . . .	17

---

*Le problème qui nous intéresse dans cette thèse est la construction de cages moléculaires. Plus particulièrement, la construction de cages moléculaires qui soient capables de reconnaître et de capturer une autre molécule, qu'on appelle aussi substrat. À partir d'un substrat donné, nous voulons construire, de manière automatique, des structures moléculaires qui puissent créer des liaisons avec ce substrat précis et l'englober afin de pouvoir le capturer.*

*La majeure partie des travaux qui ont été effectués en chimie sur les cages moléculaires prennent le problème dans le sens inverse. Pour une cage moléculaire donnée, les chimistes essaient de trouver le ou les substrats les plus compatibles. Le substrat adéquat est majoritairement trouvé à partir d'expériences et la modélisation n'est utilisée que dans le but de confirmer ces expériences. Dans la littérature informatique, le processus est similaire. Les molécules (cages et substrats) sont modélisées et à partir de ces modèles, ils recherchent quels sont ceux qui pourront interagir ensemble.*

*L'approche que nous proposons est différente. Nous voulons, à partir d'un substrat donné, modéliser des cages moléculaires spécifiques à ce substrat. Pour ce faire, nous utilisons les propriétés chimiques des atomes ainsi que des méthodes de géométrie algorithmique et d'algorithmique de graphes, pour générer des solutions.*

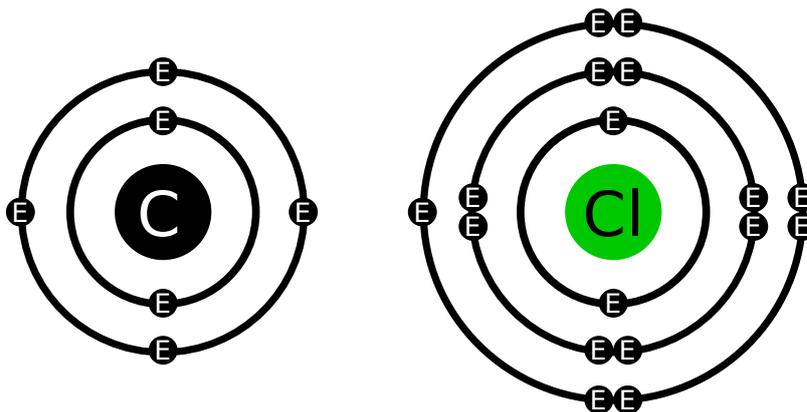
*Afin de faciliter la compréhension du sujet, nous allons replacer le contexte chimique dans lequel se place cette thèse. Nous détaillerons ensuite les travaux qui ont déjà été réalisés sur les cages moléculaires, aussi bien dans le domaine de la chimie que dans celui de l'informatique. Et nous finirons en donnant une première idée de l'approche que nous proposons ici. Dans ce chapitre nous posons également plusieurs définitions qui seront utilisées tout au long du manuscrit.*

# 1 Contexte de la thèse

Nous allons, dans cette première section, effectuer quelques rappels sur les atomes et sur les différentes liaisons qui peuvent les relier. Nous parlerons également des fondements de la chimie supra-moléculaire (science étudiant les interactions entre molécules) et plus particulièrement des cages moléculaires.

## 1.1 Rappel de chimie

La base de l'étude de la chimie est l'atome. Un atome est constitué d'un noyau central autour duquel gravitent des électrons. Comme le montre le schéma de la Figure 1.1, les électrons se répartissent par couches autour du noyau en remplissant en priorité les couches les plus proches du noyau. Ces couches sont appelées *couches électroniques*. Chaque couche électronique peut contenir un maximum de  $2n^2$  électrons avec  $n$  le numéro de la couche. La dernière couche électronique d'un atome est appelée *couche de valence*. Elle peut être partiellement ou totalement remplie. Si on reprend les exemples de la Figure 1.1, la couche de valence de l'atome de Carbone est la deuxième couche et possède quatre électrons. Celle de l'atome de Chlore est la troisième et a sept électrons.



(a) Schéma d'un atome de Carbone. (b) Schéma d'un atome de Chlore

FIGURE 1.1 – Exemples atomes et de leurs couches électroniques.

Ce sont les électrons de la couche de valence qui interviennent dans les liaisons chimiques. La liaison chimique la plus forte qu'il peut y avoir entre deux atomes est la liaison de covalence.

Afin d'être stables les électrons de la couche de valence ont besoin d'être associés à un autre électron. C'est pourquoi certains électrons interagissent avec des électrons d'autres molécules, créant ainsi des liaisons covalentes. Les autres électrons se répartissent par deux autour du noyau central.

**Definition 1.** Une *liaison covalente* est une liaison chimique dans laquelle deux atomes se partagent deux électrons de leur couche de valence afin de former un doublet d'électrons liant les deux atomes.

Il existe trois types de liaisons covalentes. Les liaisons simples qui impliquent deux électrons, les liaisons doubles qui en impliquent quatre, et les liaisons triples qui impliquent six électrons. La figure 1.2 est une représentation schématique de ces trois

types de liaisons.

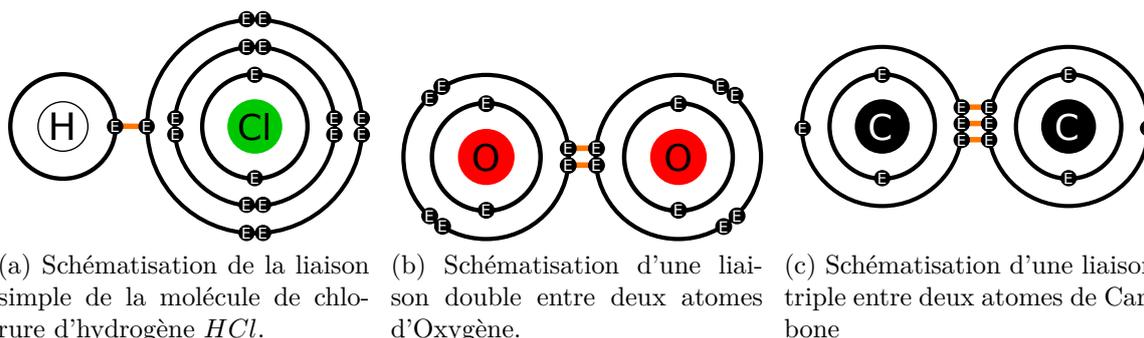


FIGURE 1.2 – Schématisation de liaisons covalentes et répartition des électrons autour des atomes liés.

Si plusieurs atomes sont reliés par des liaisons covalentes alors ils forment une molécule.

**Definition 2.** On appelle **doublets liants** les électrons de valence (appartenant à la couche de valence) qui sont impliqués dans une même liaison covalente.

Si une liaison double ou triple se forme entre deux atomes, alors le doublet liant est composé de quatre ou six électrons de la couche de valence, soient deux ou trois électrons de chaque atome.

**Definition 3.** À l'inverse, les **doublets non-liants** sont les doublets d'électrons de valence qui ne sont pas impliqués dans une liaison covalente.

Si on prend en exemple la molécule de chlorure d'hydrogène de la Figure 1.2a, l'atome de chlore a un doublet liant et trois doublets non-liants, et l'atome d'hydrogène a seulement un doublet liant.

De même, si on prend en exemple la Figure 1.2b, les atomes d'oxygène ont un doublet liant et deux doublets non-liants.

## 1.2 Chimie supra-moléculaire

La chimie supra-moléculaire est une branche de la chimie qui étudie les interactions non-covalentes, appelées aussi interactions faibles, entre les atomes. Ces interactions peuvent s'effectuer entre des atomes d'une même molécule ou entre des atomes de molécules différentes. Dans le cas des cages moléculaires, on s'intéresse aux interactions entre atomes de molécules différentes. En effet, les cages moléculaires sont des molécules qui doivent pouvoir créer des interactions faibles avec les molécules qu'elles doivent piéger.

### 1.2.1 Liaisons faibles

Il existe plusieurs types de liaisons faibles. Dans le cadre de cette thèse, on va s'intéresser plus particulièrement aux liaisons hydrogènes et aux liaisons aromatiques qui sont parmi les interactions les plus souvent rencontrées des liaisons faibles.

**Definition 4.** *Une **liaison hydrogène** est une force intermoléculaire qui implique un atome d'hydrogène et un atome électronégatif tel que l'oxygène, l'azote et le fluor.*

Pour qu'une liaison hydrogène puisse se former, un donneur et un accepteur sont nécessaires. Le donneur dans une liaison hydrogène est composé d'un atome d'hydrogène relié à un hétéroatome, c'est-à-dire un atome qui n'est ni un carbone, ni un hydrogène et qui est non métallique. Les hétéroatomes les plus fréquents sont l'oxygène, l'azote, le soufre, le phosphore et les halogènes (fluor, chlore, brome et iode). L'accepteur est un atome d'azote, d'oxygène ou de fluor possédant au moins un doublet non-liant.

La Figure 1.3 représente le schéma d'une liaison hydrogène.

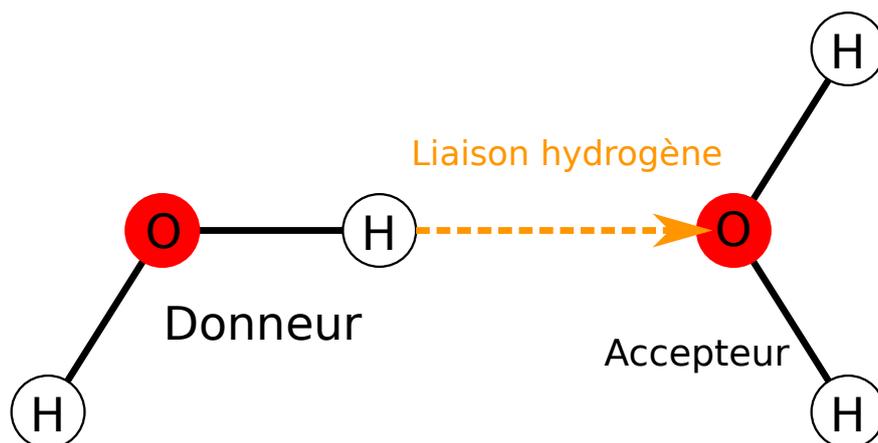


FIGURE 1.3 – Schématisation d'une liaison hydrogène entre deux molécules d'eau.

L'autre type de liaison faible qui va nous intéresser est la liaison aromatique. Les liaisons aromatiques sont des liaisons qui mettent en interaction les atomes des cycles aromatiques.

**Definition 5.** *Un **cycle moléculaire** est une série d'atomes liés de manière successive par des liaisons covalentes de sorte que les deux extrémités soient reliées.*

Dans le cadre de cette thèse, nous n'étudions que les cycles composés au maximum de six atomes.

**Definition 6.** *Un **cycle aromatique** est un cycle moléculaire dont les atomes sont situés dans un même plan.*

La particularité de ces cycles est que plusieurs des électrons se délocalisent, c'est-à-dire qu'ils ne sont plus associés à une seule liaison ou un seul atome.

Considérons, par exemple, le benzène. C'est une molécule composée de six atomes carbone et six atomes d'hydrogène. Les six carbones du benzène forment un cycle. En théorie les cycles du benzène se constituent d'une alternance de liaisons doubles et de liaisons simples comme sur la Figure 1.4a. Cependant dans la réalité les électrons se comportent comme sur la Figure 1.4b. Il y a six liaisons simples et les six électrons qui auraient dû servir aux liaisons doubles sont délocalisés et se déplacent entre les carbones du cycle. C'est ce qu'on appelle un nuage électronique aromatique.

Dans le cadre de la thèse, on va s'intéresser aux interactions entre deux cycles aromatiques. En effet, si on superpose deux cycles aromatiques, alors les électrons du nuage électronique aromatique vont interagir, créant ainsi une attraction entre les deux cycles.

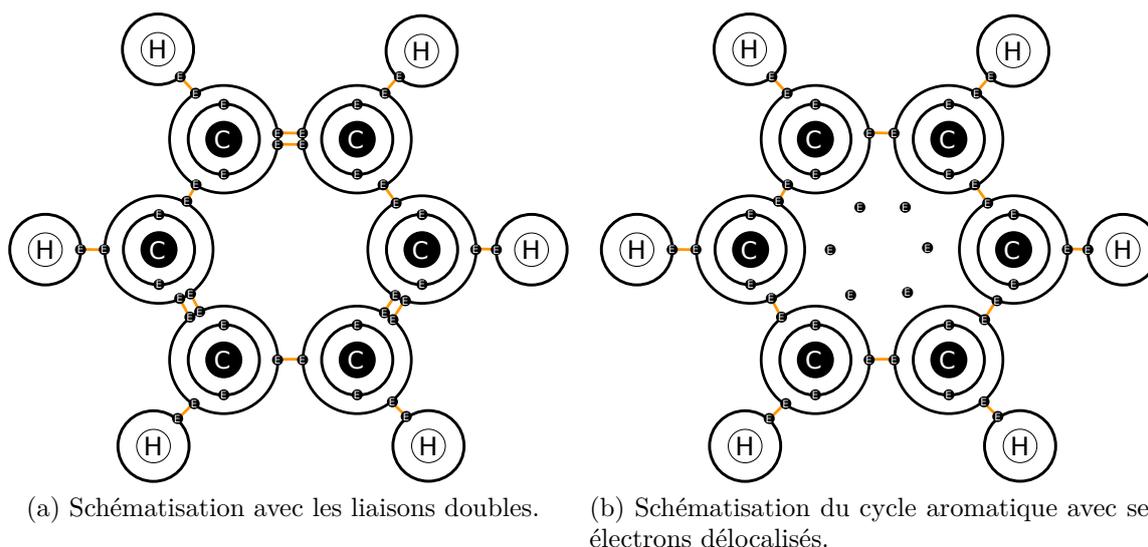


FIGURE 1.4 – Schématisation d’une molécule de benzène.

**Definition 7.** On appellera, ici, *liaison aromatique* la liaison qui peut s’établir entre deux cycles aromatiques.

### 1.3 Objectif : Cages moléculaires

Dans cette thèse, nous nous intéressons aux interactions entre molécules et plus précisément aux cages moléculaires.

**Definition 8.** Une molécule capable de reconnaître et de capturer une autre molécule, appelée substrat, est une *cage moléculaire*.

Les cages moléculaires sont des molécules qui possèdent une cavité pouvant accueillir (capturer) grâce à des interactions chimiques, une autre molécule(substrat) de forme et de taille complémentaire (reconnaissance).

Une cage moléculaire peut être totale ou partielle. Elle est dite *totale* si elle englobe entièrement le substrat. Si, à l’inverse, seule une partie de la cage est reliée au substrat, elle est *partielle*.

Pour qu’une molécule puisse être définie comme une cage moléculaire pour un substrat donné elle doit posséder deux propriétés par rapport à lui : la compatibilité et la spécificité.

**Definition 9.** Étant donné une cage et un substrat, on appelle *compatibilité* entre ces deux molécules leur capacité à pouvoir créer des interactions.

Comme expliqué précédemment, deux molécules peuvent créer des interactions faibles entre elles en formant par exemple des liaisons hydrogènes ou des liaisons aromatiques. Plus une cage peut créer de liaisons faibles « en même temps » avec le substrat plus sa compatibilité est importante. La Figure 1.5 montre un exemple de compatibilité entre une cage et un substrat. Dans cet exemple, il y a deux zones qui peuvent être utilisées pour créer des interactions entre le substrat et la cage. Dans la Figure 1.5a la distance qui sépare les deux zones de la cage est plus petite que celle qui sépare les deux zones du substrat. Ainsi les deux liaisons ne peuvent pas s’établir en même temps. La compatibilité de la cage avec le substrat est donc de 1

car ils ne peuvent établir qu'une liaison à la fois. Dans le cas de la Figure 1.5b, la distance permet aux liaisons de pouvoir s'établir en même temps, la cage a donc une compatibilité de 2 avec le substrat.

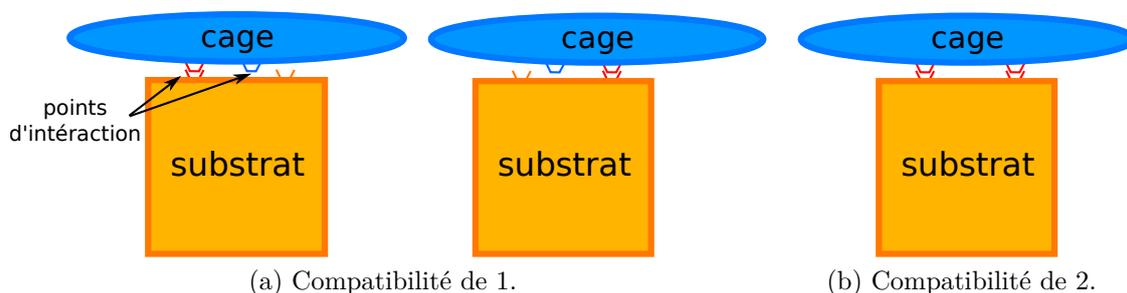


FIGURE 1.5 – Exemple de compatibilité entre une cage et un substrat.

La deuxième propriété que doit posséder la cage par rapport au substrat est la spécificité.

**Definition 10.** La *spécificité* d'une cage moléculaire avec un substrat est sa capacité à pouvoir créer uniquement des interactions avec ce substrat, c'est-à-dire sa capacité à pouvoir le reconnaître.

Plus la forme de la cage est complémentaire à celle du substrat, plus elle est spécifique au substrat. En effet, si par exemple, comme sur la Figure 1.6a, la cage est très plate, beaucoup de molécules pourront aisément créer des liaisons avec elle. À l'inverse, on voit sur la Figure 1.6b que si la forme de la cage n'est pas complémentaire avec celle du substrat, les interactions entre les deux ne pourront pas s'effectuer.

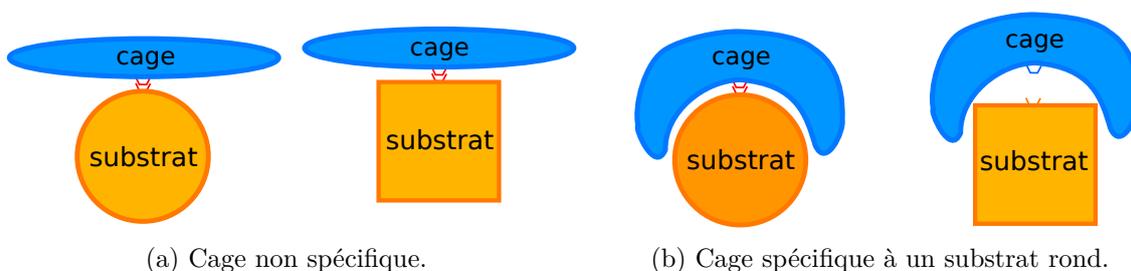


FIGURE 1.6 – Exemples de spécificité entre une cage et un substrat.

## 2 Apparition et utilisation des cages moléculaires

Comme expliqué précédemment, la chimie supra-moléculaire est l'étude des interactions intermoléculaires, c'est-à-dire des interactions entre molécules. Parmi les thématiques de la chimie supramoléculaire, l'une d'entre elles est l'interaction et la reconnaissance moléculaire dont font partie les cages moléculaires. Bien que les *cyclophanes* soient les premières structures cages à apparaître dans la littérature ([Brown and Farthing \(1949\)](#); [Cram and Steinberg \(1951\)](#)), les premières fabriquées par l'homme sont les *éthers couronnes* ([Pedersen \(1967\)](#)) qui apparaissent en 1967 grâce au chimiste

américain Charles John Pedersen et sa méthode de synthèse. Ces éthers couronnes sont capables de reconnaître et de capturer des cations. Quelques années plus tard, c'est le chimiste américain Donald James Cram, qui en utilisant la même méthode de synthèse inventée C. J. Pedersen, présente ses travaux sur les *sphérands* (Gokel and Cram (1973)), molécules avec des structures plus complexes que les éthers couronnes. En effet, alors que les éthers couronnes sont des molécules bidimensionnelles, les sphérands sont tridimensionnelles. D. J. Cram utilise la forme des molécules pour sélectionner des composés chimiques pouvant être capturés. Enfin dans les années 70, c'est au tour du français Jean-Marie Lehn de mettre en avant les *cryptands* (Lehn (1987)). Formées d'un assemblage de cycle comprenant au moins trois sites liants, ces molécules peuvent capturer des cations, des anions ou des espèces neutres.

Cependant c'est à partir de 1987, avec l'obtention du prix Nobel de Chimie de ces trois chimistes pour leurs travaux respectifs sur les éthers couronnes, les sphérands et les cryptands, que l'intérêt pour ces structures moléculaires possédant un espace intérieur défini pouvant capturer d'autres molécules va augmenter (Lehn (1999)). D'autres types de cages moléculaires vont être synthétisées au cours des années, chacune avec des propriétés différentes ce qui permet leur utilisation dans de nombreuses applications biochimiques. La synthèse de ces molécules très complexes est un intérêt de recherche à part entière. Elle repose généralement sur l'assemblage dynamique de petits modules réactifs (Mastalerz (2010)). Parmi les applications des cages moléculaires, on retrouve par exemple l'encapsulation moléculaire ou incorporation de biomolécules (bio-MOF), catalyse biomimétique, biocapteurs, libération de NO bioactif, et biomédicaments (Ahmad et al. (2015)).

L'encapsulation d'ingrédients actifs dans les molécules cages est l'une des applications les plus prometteuse (Hof et al. (2002)). Elle permet entre autre de protéger des substances contre les interactions extérieures, comme par exemple certaines molécules dans des crèmes pour éviter qu'elles irritent la peau, ou encore d'encapsuler des molécules pour les transports, comme pour le transport de médicaments (Byrne et al. (2017); Ellis-Davies and R (2007); Geldenhuys et al. (2005)). Leur nature biodégradable les prédispose à d'importantes applications dans les domaines agroalimentaire et pharmaceutique. Des applications plus avancées n'utilisent pas seulement ces structures moléculaires comme des conteneurs mais comme des enzymes artificielles utilisées pour des fonctions spécifiques (Barth et al. (2015); Mastalerz (2012); Zhang and Mastalerz (2014)).

Les recherches de ces cages moléculaires ont toutes un point commun. On a d'abord cherché à construire des cages moléculaires avec des propriétés spécifiques pour ensuite déterminer ce qu'elles étaient capables de reconnaître et de capturer. Dans cette thèse nous nous sommes intéressés au problème inverse, à savoir que à partir d'un substrat donné, nous cherchons à déterminer des topologies de cages adaptées pour le reconnaître et le capturer. Notre approche utilise des concepts tels que celles *l'empreinte moléculaire* (Mosbach (1994); Alexander et al. (2006); Whitcombe et al. (2014)) et de l'assemblage de petits modules moléculaires (Mastalerz (2010)).

### 3 Approche proposée

Étant donné un substrat cible, le but est de construire des *graphes moléculaires* modélisant des cages moléculaires capables de capturer ce substrat.

**Definition 11.** *Un graphe  $G = (V, E)$  est un **graphe moléculaire** si chaque sommet de  $V$  possède des coordonnées dans un espace 3D et représente un atome, chaque arête de  $E$  représente une liaison covalente et l'ensemble du graphe respecte trois contraintes :*

- *La distance spatiale entre chaque paire de sommets reliés par une arête doit être équivalente à la distance d'une liaison covalente.*
- *Le degré de chaque sommet ne doit pas dépasser le nombre de liaisons que peut avoir un atome.*
- *Les angles séparent les voisins d'un sommet doivent respecter la théorie de VSEPR (Gillespie (1963)).*

Les graphes moléculaires sont des structures qui permettent de connaître la position des atomes les uns par rapport aux autres, ainsi que les liaisons qu'ils partagent, mais les graphes moléculaires ne donnent pas forcément d'informations sur le type des atomes. C'est pourquoi, un graphe moléculaire pourra donner plusieurs cages moléculaires différentes en fonction des atomes utilisés.

Les atomes les plus courants en chimie organique sont les atomes de carbone, d'hydrogène, d'azote et d'oxygène. Chaque sommet des graphes moléculaires générés devra pouvoir être instancié par au moins l'un de ses quatre atomes.

Dans la suite nous détaillons les contraintes des graphes moléculaires, ainsi que ce que nous appelons des *motifs moléculaires*. Puis nous donnons une idée générale de l'approche que nous proposons.

#### 3.1 Contraintes des graphes moléculaires

Pour qu'un graphe dans lequel chaque sommet a une position dans l'espace soit un graphe moléculaire, c'est-à-dire qu'il puisse modéliser une molécule, trois contraintes doivent être respectées. La première contrainte concerne la distance spatiale qui sépare deux sommets voisins du graphe. La seconde contrainte concerne le nombre de voisins de chaque sommet du graphe. Et la dernière concerne les angles qui séparent deux voisins d'un même sommet du graphe. La partie suivant détaille ces trois contraintes.

##### 3.1.1 Distance entre deux sommets et rayon de covalence

Comme expliqué précédemment, les arêtes d'un graphe moléculaire représentent des liaisons covalentes. La distance qui sépare deux sommets voisins du graphe est donc importante pour que ce graphe modélisé soit une molécule. On définit par  $D_{[u,v]}$  la distance entre deux sommets  $u$  et  $v$  du graphe. On définit également  $D_{max}$  et  $D_{min}$  tels que quelle que soit l'arête  $[u, v]$  d'un graphe moléculaire,  $D_{min} \leq D_{[u,v]} \leq D_{max}$ . À l'inverse si  $[u, v]$  n'appartient pas au graphe alors  $D_{max} < D_{[u,v]}$ . Afin de déterminer  $D_{max}$  et  $D_{min}$ , on utilise les rayons de covalence des atomes.

**Definition 12.** *Le rayon de covalence d'un atome correspond à la moitié de la distance entre deux noyaux atomiques identiques liés par une liaison covalente simple.*

Il s'agit donc du rayon de l'atome. Le tableau 1.1 donne les rayons de covalence théoriques des quatre atomes apparaissant le plus fréquemment dans les molécules organiques, c'est-à-dire les atomes d'hydrogène, de carbone, d'azote et d'oxygène.

TABLE 1.1 – Rayons de covalence

Nom	Type	Notation	Rayon de covalence ( <i>pm</i> )
Hydrogène	H	$r_H$	30
Carbone	C	$r_C$	70
Azote	N	$r_N$	65
Oxygène	O	$r_O$	60

Étant donné que le rayon de covalence de l'atome d'Hydrogène  $r_H$  est le plus petit des rayons de covalence, on définit  $D_{min} = 2 * r_H - \epsilon$ , avec  $\epsilon$  une marge d'erreur, et  $D_{max} = 2 * r_C + \epsilon$ . La distance qui sépare deux sommets du graphe reliés par une arête doit donc être comprise entre  $2 * r_H - \epsilon$  et  $2 * r_C + \epsilon$ .

### 3.1.2 Nombre de voisins, angles et méthode de VSEPR

La méthode de VSEPR (Répulsion des Paires d'Électrons des couches de Valence), aussi appelée « théorie de Gillespie », est une méthode qui permet de déduire la géométrie d'une molécule simple, c'est-à-dire qu'elle permet de déterminer la position dans l'espace d'atomes autour d'un atome central.

Comme expliqué précédemment, un atome est constitué d'un noyau central autour duquel gravite des électrons. Ces électrons exercent des forces de répulsion les uns par rapport aux autres. Ainsi les électrons se tiendront le plus éloigné possible les uns des autres. Grâce à cette propriété des électrons, il est possible de déterminer comment les voisins d'un atome se positionneront autour de lui.

Il y a cependant une différence entre les électrons qui sont impliqués dans une liaison covalente, c'est-à-dire ceux faisant partie d'un doublet liant, et ceux n'étant pas impliqués dans une liaison covalente, c'est-à-dire ceux faisant partie d'un doublet non-liant.

Les doublets non-liants d'un atome ont une plus grande force de répulsion que les doublets liants. L'angle qui sépare deux doublets non-liants est donc plus important que l'angle qui sépare un doublet liant d'un doublet non-liant. De même, l'angle qui sépare un doublet liant d'un doublet non-liant est plus grand que l'angle qui sépare deux doublets liants. En conséquence, les angles séparant deux sommets d'un sous-graphe par rapport au sommet central sont inférieurs ou égaux à l'angle de référence de la géométrie.

Le tableau 1.2 récapitule les géométries qui peuvent être adoptées par les atomes de carbone, d'hydrogène, d'azote et d'oxygène, en fonction de leur nombre de doublets. l'atome est noté  $A$ , les doublets liants sont notés  $X$  et les doublets non-liants sont notés  $E$ . De plus,  $n$  est le nombre de doublets liants et  $m$  est le nombre de doublets non-liants.

TABLE 1.2 – Représentation de VSEPR en fonction du nombre de doublets des atomes.

Doublets	Représentation 2D	Représentation 3D	Angle	Marge
$n + m = 4$ ( <b>Géométrie tétraédrique</b> )				
$n = 4, m = 0$			$109.5^\circ$	$3^\circ$
$n = 3, m = 1$			$107^\circ$	$3^\circ$
$n = 2, m = 2$			$109.28^\circ$	$3^\circ$
$n = 1, m = 3$			$109.28^\circ$	$3^\circ$
$n + m = 3$ ( <b>Géométrie triangulaire</b> )				
$n = 3, m = 0$			$120^\circ$	$2^\circ$
$n = 2, m = 1$			$120^\circ$	$2^\circ$
$n = 1, m = 2$			$120^\circ$	$2^\circ$
$n + m = 2$ ( <b>Géométrie linéaire</b> )				
$n = 2, m = 0$			$180^\circ$	-
$n = 1, m = 1$			$180^\circ$	-
$n + m = 1$				
$n = 1, m = 0$			-	-

Les atomes de carbone, d'hydrogène, d'azote et d'oxygène peuvent avoir entre 1 et 4 voisins. Cette propriété est également vraie pour une grande majorité des atomes qui ont besoin de quatre doublets pour combler leur couche électronique extérieure et ainsi être stables. On limite donc à quatre le degré des sommets d'un graphe moléculaire.

De même, la méthode de VSEPR nous permet également de définir une contrainte sur les angles séparant deux voisins d'un même sommet en fonction de son degré. On a trois géométries possibles : linéaire, triangulaire ou tétraédrique. Les angles séparant les voisins d'un sommet peuvent être de  $180^\circ$  pour la géométrie linéaire, de  $120^\circ$  pour la triangulaire et de  $109.28^\circ$  pour la tétraédrique. Cependant ces angles peuvent varier légèrement en fonction du nombre de doublets liants et non-liants présents autour de l'atome, c'est pourquoi on relâche un peu cette contrainte en laissant une marge de

4 à 5°. Les angles ont également une influence sur le nombre de voisins maximum de chaque sommet. On limite à deux le nombre de voisins d'un sommet avec des angles de 180°, à trois pour les angles de 120° et à quatre pour les angles de 109.28°. Les sommets qui atteignent le nombre maximum de voisins sont appelés *sommets saturés*.

**Definition 13.** *Un sommet est dit **saturé** s'il a atteint le nombre de sommets maximum qu'il peut posséder en adéquation avec la théorie de VSEPR et le type de sommet qu'il représente, c'est-à-dire que tous ses doublets sont des doublets liants.*

## 3.2 Motifs moléculaires

Afin de définir la notion de motif moléculaire, nous commençons par définir ce que sont les sommets *saturés* et les sommets *insaturés*.

**Definition 14.** *Un sommet est dit **insaturé** s'il n'a pas atteint le nombre de sommets maximum qu'il peut posséder en adéquation avec la théorie de VSEPR.*

**Definition 15.** *Un **motif moléculaire** est un graphe moléculaire qui possède des sommets saturés et insaturés.*

Puisque les motifs moléculaires possèdent des sommets insaturés, ils peuvent être reliés entre eux grâce à ces sommets jusqu'à saturation. On définit deux types de motifs moléculaires dans cette thèse.

Pour commencer, il y a les motifs hydrogènes. Un *motif hydrogène* est un motif moléculaire représentant un ensemble d'atomes pouvant être impliqués dans une liaison hydrogène. Pour qu'une liaison hydrogène puisse se former, il faut un ensemble d'atomes donneur et un ensemble d'atomes accepteur. Les motifs hydrogènes représentent ces ensembles d'atomes spécifiques aux liaisons hydrogènes. On aura donc des motifs hydrogènes donneurs et des motifs hydrogènes accepteurs. Si un motif hydrogène donneur est présent dans le modèle substrat, un motif accepteur pourra être présent dans le graphe moléculaire de la cage. De même, si un motif accepteur est présent dans le substrat, un motif donneur pourra apparaître dans le graphe de la cage. On dit que les motifs donneurs sont complémentaires aux motifs accepteurs et inversement.

Les deuxièmes types de motifs sont les motifs aromatiques. Ce sont des motifs représentant des cycles moléculaires. Si un motif aromatique est présent dans le substrat, alors on placera un motif aromatique au même niveau dans la cage.

Ces deux types de motifs moléculaires sont décrits comme « liants » car ils représentent des ensembles d'atomes pouvant être impliqués dans des liaisons. Ces différents motifs seront détaillés dans le chapitre 3.

## 3.3 Idée générale

L'approche que nous proposons ici est de générer des guides qui serviront de modèles pour construire des cages moléculaires capables de reconnaître et de capturer un substrat donné.

L'idée de l'approche est de passer par une structure intermédiaire que nous appelons **enveloppe**. Cette enveloppe est ensuite utilisée pour générer des guides pour la construction des cages moléculaires. L'enveloppe est construite de manière à être très spécifique au substrat donné. Bien qu'il ne s'agisse pas d'un graphe moléculaire, puisqu'elle ne remplit pas les contraintes énoncées précédemment, elle est malgré tout modélisée par un graphe dont les sommets ont des coordonnées dans l'espace.

L'enveloppe est une structure qui englobe le substrat et sa forme générale est complémentaire à celle du substrat. La forme de l'enveloppe peut être considérée comme un moulage de celle du substrat à la différence qu'un espace est laissé entre le substrat et l'enveloppe. Cet espace est approximativement celui d'une liaison faible pouvant exister entre deux molécules. Ainsi l'enveloppe est utilisée pour délimiter une zone à l'intérieur de laquelle les sommets des cages générées ne pourront pas se trouver.

Le deuxième intérêt de cette enveloppe est qu'elle est construite dans une logique moléculaire. Les sommets sont positionnés de telle sorte qu'ils soient à des distances et dans des directions dans lesquelles les atomes du substrat peuvent créer des interactions (des liaisons) avec des atomes d'autres molécules.

**Definition 16.** *On définit par **directions libres** des vecteurs allant dans la direction et le sens dans lesquels les atomes du substrat peuvent encore créer des liaisons avec d'autres atomes, c'est-à-dire les directions des doublets non-liants des atomes.*

**Definition 17.** *On définit par **position libre** une position dans l'espace qui est située dans une direction libre de l'un des sommets de  $G$  et à une distance équivalente à celle d'une liaison faible de ce sommet.*

Pour chaque atome du substrat, on va chercher l'ensemble de ses positions libres. Pour chaque position libre, un sommet est créé dans l'enveloppe. Cette étape est appelée **l'expansion du substrat**. Puis, on construit les arêtes de l'enveloppe en construisant une enveloppe concave à partir à partir de l'ensemble de ses sommets.

La phase suivante consiste à construire les guides des cages moléculaires à partir de cette enveloppe. Puisque les sommets de l'enveloppe ont été positionnés à des endroits stratégiques pour créer des interactions avec le substrat, des motifs moléculaires sont insérés dans l'enveloppe en remplaçant des sommets ou des ensembles de sommets de l'enveloppe.

Lors de cette phase, nous insérons les motifs moléculaires liants, c'est-à-dire les motifs aromatiques et hydrogènes, en face des zones où les atomes du substrat peuvent créer des interactions faibles.

La Figure 1.7 résume les différentes phases de l'approche proposée. La première phase, la construction de l'enveloppe, est divisée en trois étapes principales : la recherche de la topologie des sommets du substrat, l'extension des sommets et enfin le calcul de l'enveloppe concave à partir des extensions. Ces étapes sont approfondies dans le chapitre 2. La seconde phase est l'insertion des motifs liants dans l'enveloppe qui se fait en deux étapes (décrites dans le chapitre 3) puisque que deux motifs liants différents sont insérés. Enfin la troisième étape est la génération de graphes moléculaires représentant des molécules cages à partir des enveloppes incluant les motifs liants.

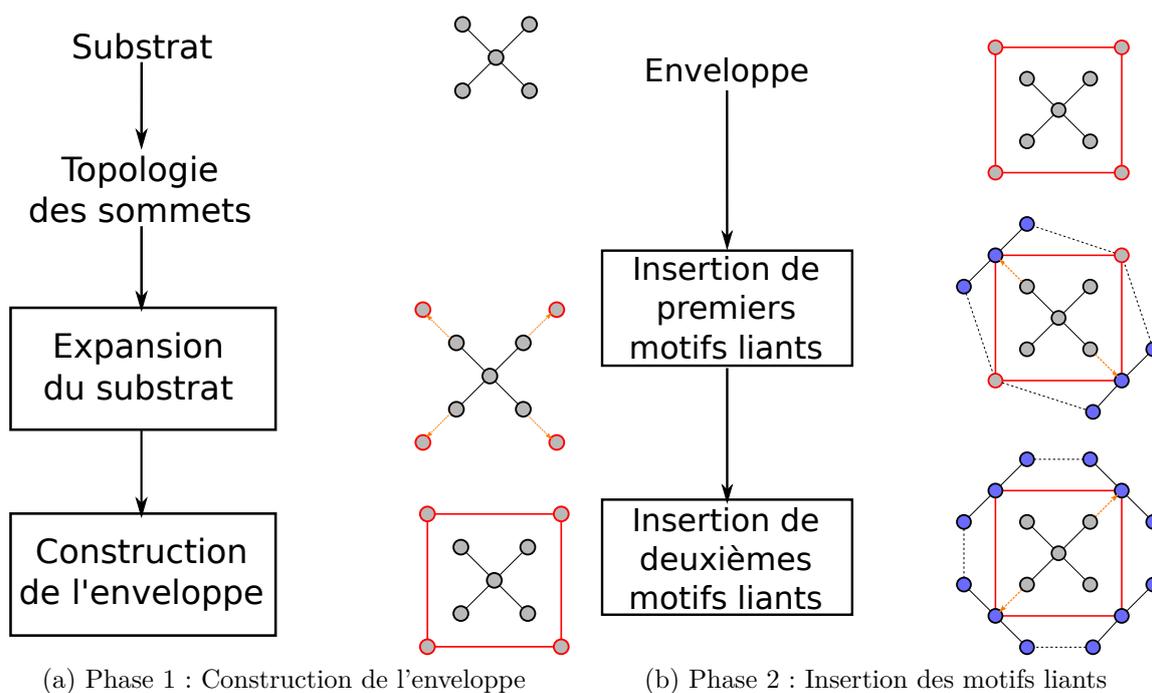


FIGURE 1.7 – Récapitulatif des phases de la génération de guide par les cages moléculaires d'un substrat donné.

## 4 Conclusion

Le problème de cages organiques en chimie supra-moléculaire, consiste à associer une molécule cible, appelée substrat à une autre molécule qui soit capable de capturer la première. En chimie, les solutions avancées pour résoudre ce problème, sont de partir d'une cage moléculaire donnée et de trouver, par des expériences, le substrat ou les substrats qu'elle est le plus apte à capturer. Cette approche est aussi la base des méthodes informatiques qui cherchent dans un premier temps à construire des cages, pour ensuite déterminer les molécules qu'elles peuvent capturer. Ces cages sont souvent très symétriques et peuvent capturer un grand nombre de substrats différents.

L'approche que nous proposons consiste, à l'inverse des méthodes déjà proposées, à partir d'un substrat cible et à générer des guides pour construire les cages les plus spécifiques à ce substrat. L'idée est de construire une première structure spécifique autour du substrat et de placer des morceaux de molécule (motifs moléculaires) dans cette structure de telle sorte que ces morceaux soient des points d'interaction entre les cages et le substrat.



# Chapitre 2

## Construction de l'enveloppe

---

1	Données et Modélisation . . . . .	20
2	Topologie des sommets de $G$ . . . . .	21
3	Expansion du substrat . . . . .	30
4	Construction des arêtes de l'enveloppe . . . . .	38
5	Conclusion . . . . .	41

**3 Intégration des motifs liants** **43**

---

*La première étape de l'approche est la construction d'une structure spécifique autour du substrat. Cette structure est appelée **enveloppe**. L'enveloppe n'est pas un graphe moléculaire mais elle est construite dans une logique moléculaire. C'est une structure qui englobe géométriquement le substrat et dont les sommets sont positionnés dans des directions où les atomes du substrat peuvent encore avoir des interactions avec d'autres atomes. Ces directions sont donc des directions dans lesquelles le substrat peut avoir des interactions avec d'autres molécules. Elles sont appelées **directions libres**. La première étape de la construction de cette enveloppe est de déterminer les directions libres des sommets du substrat en nous appuyant sur des données chimiques ainsi que la théorie de VSEPR. Chaque direction libre trouvée est utilisée pour déterminer une position libre qui sert de coordonnées spatiales à un sommet de l'enveloppe. Enfin à partir de l'ensemble des sommets nous construisons une enveloppe concave pour déterminer les arêtes de l'enveloppe afin que l'enveloppe englobe le substrat.*

*Dans ce chapitre nous détaillons la première phase de la génération des cages moléculaires, c'est-à-dire la construction de l'enveloppe. Pour ce faire, nous commençons par décrire plus précisément ce qu'est l'enveloppe et les propriétés qu'elle doit posséder afin de pouvoir être utilisée pour la construction de guides des cages moléculaires. Puis, nous expliquons les trois étapes de cette phase. Dans un premier temps nous détaillons la démarche que nous avons mise en place afin de trouver la topologie des sommets du substrat, que nous définirons après. Nous montrerons ensuite comment à partir de cette topologie nous retrouvons les directions libres des sommets du substrat afin de trouver les coordonnées (positions libres) des sommets de l'enveloppe. Et enfin nous finirons en expliquant comment nous construisons l'ensemble des arêtes de l'enveloppe à partir de la position des sommets, afin que le substrat soit à l'intérieur de la structure définie par l'enveloppe.*

# 1 Données et Modélisation

## 1.1 Données d'entrée

Il existe différents formats de données permettant de stocker les informations des molécules. Les informations stockées sont plus ou moins complètes en fonction du format utilisé. Par exemple, les formats comme *XYZ* ([Wikipédia \(a\)](#)) ou *GAMESS* ([research group at Iowa State University](#)) stockent uniquement des informations sur les atomes (type, position dans l'espace, ...) alors que d'autres formats comme *PDB* ([wwPDB Foundation](#)) ou *MDL SDfile* ([Autodesk](#)) stockent également des informations sur les liaisons entre les atomes. Afin de ne pas être tributaire d'un éventuel manque d'informations pour certaines molécules, nous avons décidé d'utiliser les informations communes minimales contenues dans n'importe quel format, à savoir : le nombre d'atomes, le type de chaque atome et leurs coordonnées dans l'espace.

Ainsi nous modélisons le substrat comme un graphe non-orienté  $G = (V_G, E_G)$ , tel que  $V_G$  est l'ensemble des atomes du substrat et  $E_G$  l'ensemble des liaisons entre les atomes. De plus, à chaque sommet de  $V_G$ , on associe également ses coordonnées dans l'espace.

L'ensemble  $V_G$  est construit à partir de la liste des atomes du substrat. Il nous faut ensuite construire l'ensemble  $E_G$  à partir des informations dans les éléments de  $V_G$ . Pour cela, on utilise les rayons de covalence des atomes (Chapitre 1, section 12). Deux atomes sont liés par une liaison covalente si leur couche de valence sont en contact. On connaît les valeurs approximatives des rayons de covalence des atomes puisqu'ils dépendent du type des atomes. On note  $r_v$  le rayon de covalence d'un sommet  $v$  de  $V_G$ . Dans la mesure où les rayons de covalence sont des moyennes, on utilise un pourcentage d'erreur de 20%.

**Definition 18.**  $\forall u \in V_G, \forall v \in V_G, \exists (u, v) \in E_G$  ssi  $(r_u \pm 20\% + r_v \pm 20\%) \geq \text{distance}(u, v)$ .

## 1.2 Enveloppe

Un graphe est une enveloppe  $S = (V_S, E_S)$  s'il respecte trois propriétés :

**Propriété 1 :** Chaque sommet de  $S$  doit être situé dans l'une des *directions libres* de l'un des sommets de  $G$ .

Chaque sommet de  $S$  est dans une des directions stratégiques pour favoriser les interactions entre le substrat et les molécules qui seront générées à partir de l'enveloppe.

**Propriété 2 :** Chaque sommet de  $S$  doit être positionné sur une position libre de  $G$ , c'est-à-dire à une distance équivalente à celle d'une liaison faible dans l'une des directions libres d'un des sommets de  $G$ .

Comme expliqué dans le chapitre précédent, lorsque deux molécules interagissent entre elles, elles forment des liaisons faibles entre certains de leurs atomes. Si les sommets de l'enveloppe sont positionnés à une distance équivalente à celle d'une liaison faible, la possibilité d'interaction entre les molécules qui seront générées et le substrat augmente. Il existe plusieurs types de liaisons faibles, c'est pourquoi on établit cette distance en utilisant comme référence la liaison faible la plus énergétique et courte qui

est la liaison hydrogène. Chaque sommet de  $V_S$  doit être situé à environ  $1.8\text{pm}$  de l'un des sommets de  $G$  qui est la distance moyenne séparant deux atomes impliqués dans une liaison hydrogène.

**Propriété 3** : Aucune des arêtes de  $S$  ne doit traverser géométriquement la structure définie par le graphe  $G$ .

Le graphe  $S$  doit être construit en respectant ces trois propriétés pour qu'il puisse efficacement servir comme fondation pour les générations des molécules cages du substrat.

**Definition 19.** On appelle *extension* d'un sommet  $v \in G$  l'ensemble de ses positions libres.

**Definition 20.** On appelle *expansion* du substrat l'ensemble des extensions des sommets de  $G$ .

Pour trouver les positions libres, il faut commencer par déterminer les directions libres des sommets de  $G$ . On a vu dans le chapitre précédent qu'il y a une relation entre le nombre de doublets d'un atome et leur répartition autour de celui-ci, c'est-à-dire sa géométrie. Dans un premier temps, on cherche donc à déterminer la géométrie de chaque sommet du graphe  $G$  afin de connaître le nombre de doublets qu'il possède et les directions dans lesquelles ces doublets sont situés.

## 2 Topologie des sommets de $G$

La première étape de la construction de l'enveloppe est la recherche de la topologie des sommets. Pour chaque sommet  $v \in V_G$ , on définit sa **topologie** comme un couple  $(n_v, m_v)$  tel que  $n_v$  est le nombre de doublets liants de  $v$  et  $m_v$  son nombre de doublets non-liants. En chimie organique,  $n_v + m_v \leq 4$ . Comme il n'y a pas de distinction entre les différents types de liaisons (simple, double, triple), le nombre de doublets liants de chaque sommet  $v$  est défini comme étant le degré de  $v$  dans  $G$ . La difficulté est donc de déterminer le nombre de doublets non-liants de chaque sommet. On distingue deux cas : le cas où le sommet de  $G$  possède au moins deux voisins et le cas où il n'en possède qu'un seul.

### 2.1 Topologie des sommets ayant deux voisins ou plus

Nous cherchons ici à déterminer la topologie des sommets possédant au moins deux voisins dans le graphe représentant le substrat. La méthode de VSEPR met en avant la relation entre le nombre de doublets d'un atome et sa géométrie dans l'espace (positions relatives de ses voisins en fonction de leur nombre). On distingue trois types de géométrie : tétraédrique, triangulaire et linéaire. Pour chaque sommet du graphe  $G$ , on cherche à déterminer laquelle de ces trois géométries il possède pour en déduire son nombre de doublets. Dans la mesure où un sommet possède au moins deux voisins, il est possible de déterminer l'angle qui les sépare. La moyenne des angles d'un sommet peut nous permettre d'en déduire sa géométrie. Cependant dans le cas où un sommet appartient à un cycle, au sens d'un cycle chimique, il peut arriver que les angles soient distordus, par rapport à l'angle théorique, par l'influence des autres sommets de ce cycle. Parmi les sommets de  $G$  possédant deux voisins, on distingue ceux appartenant à un cycle des autres sommets afin de déterminer leur géométrie et ainsi leur nombre

de doublets.

Pour trouver la topologie des sommets possédant au moins deux voisins, nous commençons par déterminer ceux qui appartiennent à des cycles. Dans un graphe moléculaire un cycle est un cycle induit de longueur maximum six. Nous étudions ensuite pour chaque sommet les angles qui séparent leurs voisins afin d'avoir une première indication sur leur géométrie théorique. Enfin, dans la mesure où les angles peuvent être distordus par rapport aux angles théoriques, nous distinguons les sommets n'appartenant pas à des cycles des autres pour déterminer leur topologie réelle.

### 2.1.1 Détermination des sommets appartenant à un cycle

On cherche à connaître la liste des sommets de  $G$  qui appartiennent à un cycle moléculaire. Dans un premier temps, on essaye de réduire au maximum la liste des sommets pouvant appartenir à un cycle. Ensuite, pour chacun des sommets restants, on vérifie si oui ou non il appartient à un cycle, comme nous le montre l'algorithme 1.

---

#### Algorithme 1 : SommetsAppartenantCycles

---

**Entrées :** Graphe du substrat  $G$

**Sorties :** Liste  $l$  les sommets appartenant à un cycle

```

1 Graphe  $T = ReductionGraphe(G)$ 
2  $l = \{\emptyset\}$ 
3 pour tous les sommets  $v$  de  $T$  faire
4   | si  $RechercheCycle(T, \{\emptyset\}, v, -1)$  alors
5   |   |  $l = l \cup v$ 
6   | fin
7 fin
8 retourner  $l$ 
```

---

Pour limiter le nombre de sommets à regarder, on commence par créer une copie du graphe dans laquelle on supprime les sommets ne pouvant pas faire partie d'un cycle. Si un sommet appartient à un cycle, c'est qu'il possède au moins deux voisins. On va donc retirer de manière successive les sommets ne possédant qu'un seul voisin. Comme on le voit dans l'algorithme 2, on commence par faire la liste des sommets n'ayant qu'un seul voisin. Après suppression d'un des sommets de cette liste, on vérifie si le voisin auquel il était rattaché doit, à son tour, être inséré dans la liste des sommets ne possédant qu'un unique voisin.

**Algorithme 2** : ReductionGraphe

---

**Entrées** : Graphe du substrat  $G$   
**Sorties** : Graphe  $R$  réduction de  $G$

- 1 Graphe  $R = copie(G)$
- 2 Liste  $l = \{\emptyset\}$
- 3 **pour tous les sommets  $v$  de  $R$  faire**
- 4     **si** le nombre de voisin de  $v$  est égal à 1 **alors**
- 5          $l = l \cup v$
- 6     **fin**
- 7 **fin**
- 8 **pour tous les sommets  $v$  de  $l$  faire**
- 9      $w = voisin\ de\ v$
- 10     $R = R \setminus \{v\}$
- 11    **si** le nombre de voisin de  $w$  est égal à 1 **alors**
- 12      $l = l \cup w$
- 13    **fin**
- 14 **fin**
- 15 **retourner  $T$**

---

La Figure 2.1 est un exemple d'un graphe et de son graphe réduit. Comme on le voit sur cette figure, il est possible que certains sommets n'appartenant à aucun cycle puissent encore être présents dans le graphe réduit. C'est pourquoi une seconde étape est nécessaire pour trouver les sommets des cycles.

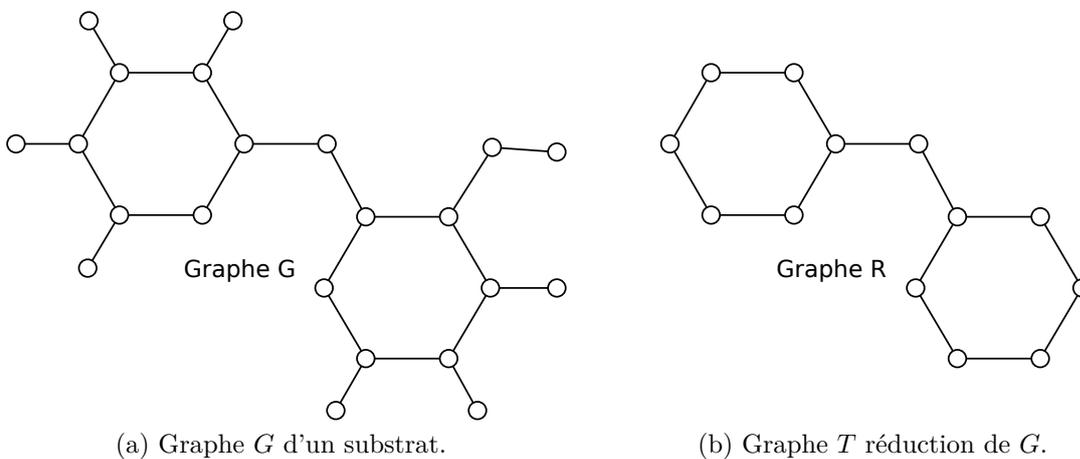


FIGURE 2.1 – Exemple de réduction d'un graphe.

Pour chaque sommet  $v$  du graphe  $R$ , on s'intéresse aux chaînes dont il est au moins l'une des extrémités. Dans la mesure où les cycles qu'on étudie ont une longueur maximum de 6, on ne s'intéresse qu'aux chaînes de taille maximum 6. À chaque itération de l'algorithme 3, on vérifie si le sommet courant est le même que le premier de la chaîne. Si ce n'est pas le cas on vérifie que la chaîne n'a pas atteint la taille maximum fixée et qu'il n'y a pas un autre cycle plus petit à l'intérieur de la chaîne. Enfin on insère le sommet courant à la chaîne et on recommence la démarche sur ses voisins.

**Algorithme 3** : RechercheCycle

---

**Entrées** : Graphe  $R$ , Liste  $l$  des sommets déjà visités, Sommet courant  $v$ ,  
Sommet précédent  $p$

**Sorties** : Booléen

```

1 Booléen  $out = FAUX$ 
2 si  $v = premier(l)$  alors
3 |   retourner  $VRAI$ 
4 fin
5 si  $Taille(l) > 5$  et  $v \in l$  alors
6 |   retourner  $FAUX$ 
7 fin
8  $l = l \cup v$ 
9 pour tous les voisins  $w$  de  $v$  faire
10 |   si  $out = FAUX$  et  $w \neq p$  alors
11 |   |    $out = RechercheCycle(R, l, w, v)$ 
12 |   fin
13 fin
14  $l = l \setminus \{v\}$ 
15 retourner  $out$ 

```

---

Les Figures 2.2 et 2.3 montrent les chemins possibles entre deux des sommets de l'exemple précédent. La Figure 2.2 montre l'exemple de l'un des sommets appartenant effectivement à un cycle. Quatre chemins peuvent être visités à partir de ce sommet. Sur les quatre, deux d'entre eux donnent une réponse positive et les deux autres une réponse négative puisque après six itérations aucun cycle n'est détecté. Cependant dans cet exemple tous les chemins ne seront pas visités car l'algorithme s'arrête à la première réponse positive.

Dans l'exemple de la Figure 2.3, on observe que le sommet testé donne aussi lieu à quatre chemins mais aucun d'entre eux n'aboutit à un cycle.

Afin de déterminer l'ensemble des sommets de  $G$  appartenant à au moins un cycle, on applique cet algorithme 3 sur tous les sommets du graphe réduit  $R$ .

### 2.1.2 Sous-graphe et angles

La géométrie d'un atome est donnée par la manière dont ses voisins sont répartis autour de lui. Pour chaque sommet de  $G$  possédant au moins deux voisins, on veut aboutir à un graphe plus simplifié ne comprenant que le sommet en question et ses voisins.

Pour chaque sommet  $v$  de  $G$  tel que  $n_v \geq 2$ , on définit  $G_v = (V_{G_v}, E_{G_v})$  comme le sous-graphe de  $G$  induit par  $v$  et ses voisins. Pour chacun de ces sous-graphes, on cherche à déterminer sa géométrie afin de connaître le nombre de doublets du sommet central. La Figure 2.4 montre des exemples de sous-graphes construits sur l'exemple précédent. Chaque sous-graphe est représenté en rouge dans cette figure.

**Definition 21.** Pour chaque couple de sommets  $(a, b)$  avec  $a$  et  $b \in V_{G_v}$  et tel que  $a$  et  $b$  sont différents de  $v$ , on définit  $\alpha_{(a,b)}$  comme la valeur de l'angle  $\widehat{avb}$ .

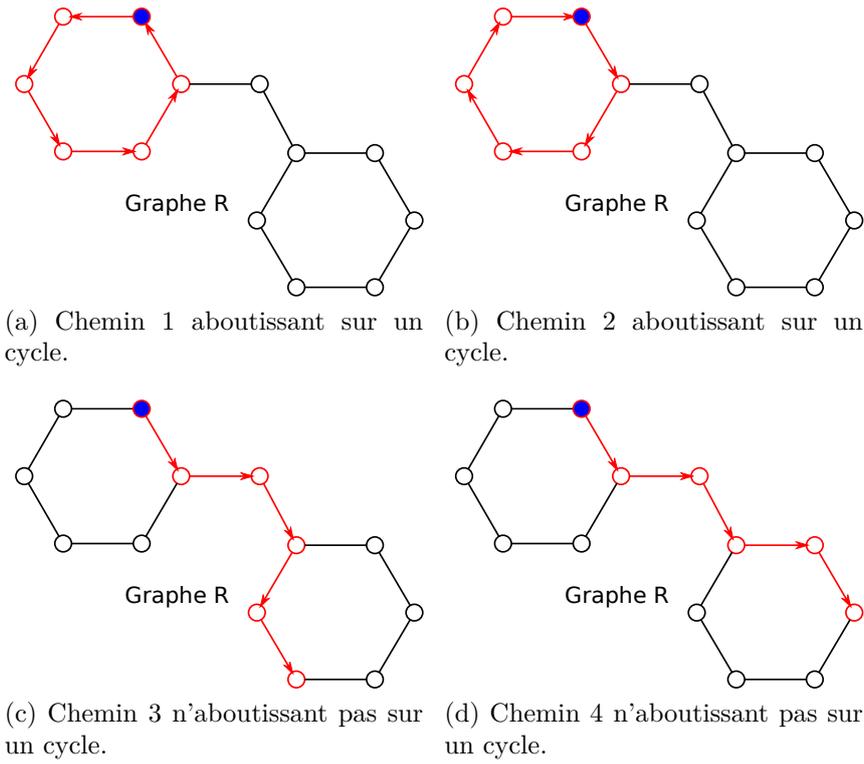


FIGURE 2.2 – Chemins trouvés pour un sommet du graphe  $R$  appartenant à un cycle.

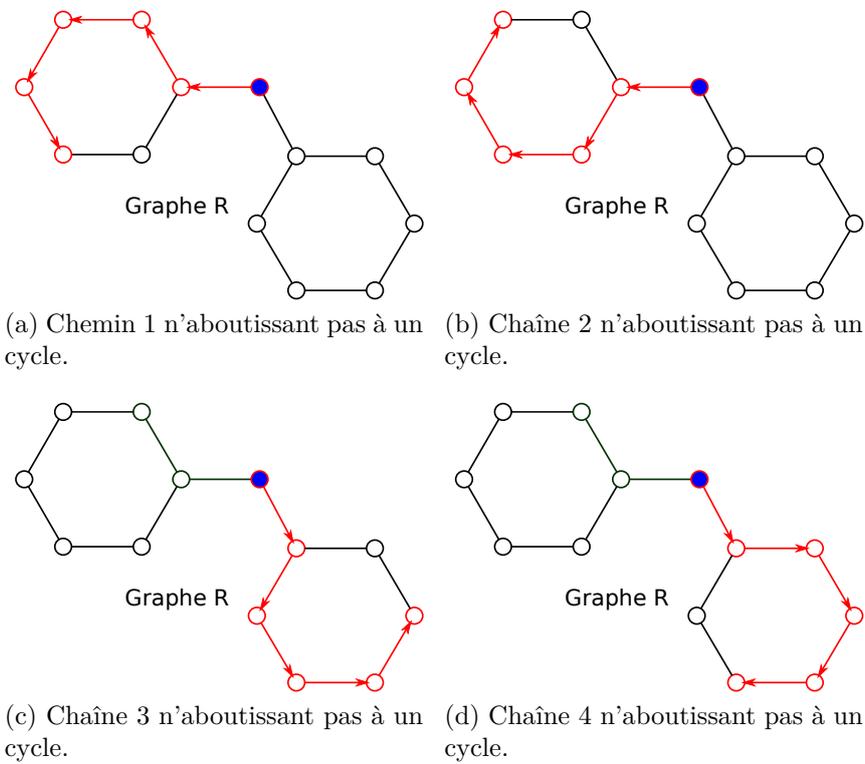
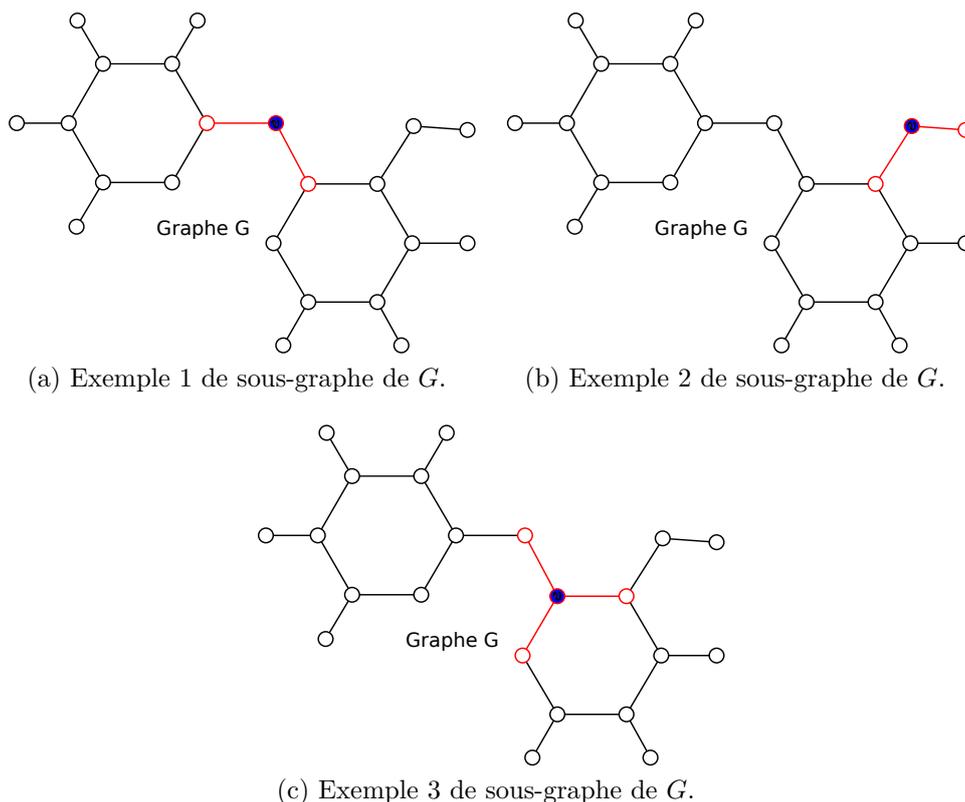


FIGURE 2.3 – Chemins trouvés pour un sommet du graphe  $R$  n'appartenant pas à un cycle.

FIGURE 2.4 – Exemples de sous-graphe de  $G$ 

**Definition 22.** On définit  $\bar{x}_v$  comme la moyenne des  $\alpha_{(a,b)}$  de  $G_v$ .

### 2.1.3 Cas général : Topologie d'un sommet n'appartenant pas à un cycle

Si le sommet n'appartient pas à un cycle, on peut se ramener directement aux résultats de la méthode de VSEPR. Pour chaque sommet, on cherche à déterminer si sa géométrie est tétraédrique, triangulaire ou linéaire afin d'obtenir sa topologie.

Avec les technologies actuelles les positions des atomes ne sont pas toujours précises, en particulier les positions des hydrogènes. On laisse donc une marge d'erreur possible. Dans chacun de ces cas,  $\epsilon$  est la marge d'erreur.

**Cas 1 :** Dans le cas où  $\bar{x}_v \leq 109.28 + \epsilon$ , on est en présence d'un sommet à géométrie tétraédrique qui possède quatre doublets, c'est-à-dire que  $n_v + m_v = 4$ . Comme le montre la Figure 2.5, trois topologies sont possibles en fonction de la valeur de  $n_v$ .

**Cas 2 :** Dans le cas où  $109.28 + \epsilon < \bar{x}_v \leq 120 + \epsilon$ , on est en présence d'un sommet à géométrie triangulaire qui possède trois doublets, c'est-à-dire que  $n_v + m_v = 3$ . Comme le montre la Figure 2.6, deux topologies sont possibles en fonction de  $n_v$ .

**Cas 3 :** Dans le cas où  $180 - \epsilon \leq \bar{x}_v \leq 180 + \epsilon$ , on est en présence d'un sommet à géométrie linéaire qui possède deux doublets. Sa topologie est alors de type  $(2, 0)$  (voir Figure 2.7).

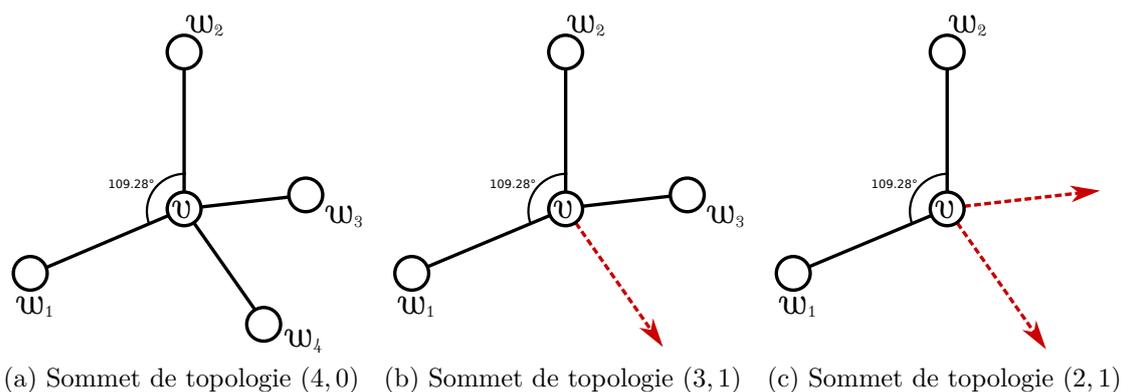


FIGURE 2.5 – Sommets à géométrie tétraédrique

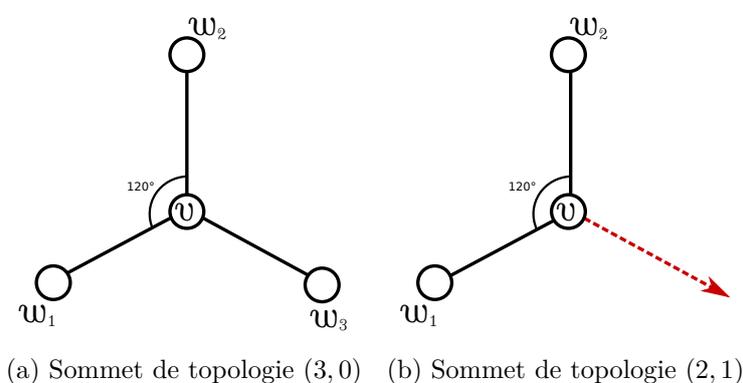


FIGURE 2.6 – Sommets à géométrie triangulaire

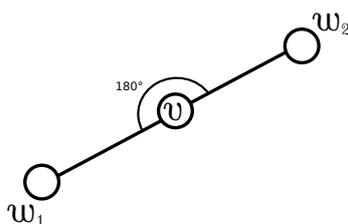


FIGURE 2.7 – Sommet à géométrie linéaire de topologie (2, 0)

### 2.1.4 Cas particulier : Topologie d'un sommet appartenant à un cycle

Dans un cycle, il peut arriver que les angles théoriques soient déformés. C'est par exemple le cas des cycles à 5 atomes. Comme le montre la Figure 2.8, dans un cycle à 5 sommets (pentagone régulier) les angles sont en moyenne de  $108^\circ$ . Tous les atomes du cycle sont dans un même plan et leur topologie est de type (2, 1) malgré l'angle de  $108^\circ$  au lieu de  $120^\circ$ . C'est pourquoi on ne peut pas uniquement se fier aux angles dans le cas des cycles.

En chimie, les cycles sont réguliers, c'est-à-dire que l'ensemble des atomes qui

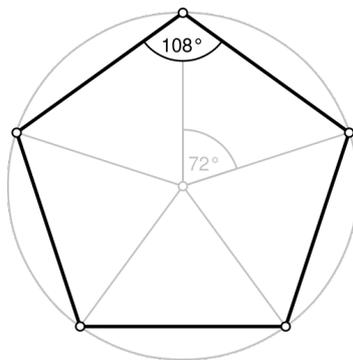


FIGURE 2.8 – Cycle régulier à 5 sommets.

les composent ont la même géométrie. Si on trouve la géométrie d'un des sommets du cycle, on peut en déduire la géométrie des autres sommets du même cycle. Pour commencer, on cherche donc les sommets qui ne possèdent pas de doublets non-liants. Si un sommet  $v$  possède 4 voisins, il est alors de topologie  $(4, 0)$ . De même si un sommet  $v$  possède 3 voisins et que  $\bar{x}_v = 120$ , alors il est de topologie  $(3, 0)$ . Pour les autres sommets du cycle, on propage le résultat. Si un sommet  $v$  appartient à un cycle et qu'il possède au moins un voisin avec 3 doublets, alors  $v$  a une topologie de  $(2, 1)$  sinon  $v$  a 4 doublets.

## 2.2 Topologie des sommets possédant un voisin unique

Dans le cas où un sommet ne possède qu'un seul voisin, on ne peut pas déterminer sa géométrie. Par conséquent, il n'est pas possible d'utiliser la méthode précédente. C'est pourquoi d'autres règles sont établies. On distingue deux cas. Le premier est le cas où le sommet représente un atome ne pouvant avoir qu'un seul doublet liant de par sa nature chimique (les hydrogènes ou les halogènes), alors la topologie du sommet est directement dépendante du type de l'atome qu'il représente. Dans les autres cas, on utilise la topologie du voisin du sommet car elle a une influence sur la topologie du sommet lui-même. On établit donc une corrélation entre le nombre de doublets du sommet et celui de son voisin.

### 2.2.1 Cas particulier : les hydrogènes

Les atomes d'hydrogène ne possèdent qu'un seul doublet et n'ont pas de doublets non-liants. En effet, ils ne possèdent qu'un seul électron, pour saturer leur couche externe et ainsi être stables, ils ont besoin d'un autre électron (annexe A, règle du duet) qu'ils obtiennent en formant une liaison avec un autre atome. Les atomes d'hydrogène devraient donc avoir une topologie de type  $(1, 0)$ . Cependant ils peuvent être impliqués dans des liaisons hydrogènes, c'est pourquoi on les référence dans le modèle comme des sommets de type  $(1, 1)$ .

### 2.2.2 Cas particulier : les halogènes ( $F$ , $Cl$ , $Br$ , $I$ )

Le fluor ( $F$ ), le chlore ( $Cl$ ), le brome ( $Br$ ) et l'iode ( $I$ ) sont des atomes qui font partie de la famille des halogènes. Leur particularité est qu'ils n'ont besoin que d'un seul électron pour saturer leur couche externe (annexe A, règle de l'octet). Ils ont donc toujours un seul doublet liant et 3 doublets non-liants. Les sommets des atomes qui les représentent dans le modèle sont donc toujours de topologie  $(1, 3)$ .

### 2.2.3 Cas général des sommets ayant un seul voisin

La topologie des autres sommets ne possédant qu'un seul voisin dépend de la liaison qui les relie à leur voisin. En effet, s'ils sont reliés par une liaison triple, il ne reste donc qu'un seul doublet non-liant possible et le sommet est alors de topologie  $(1, 1)$ . S'ils sont reliés par une liaison double, le sommet a alors deux doublets non-liants et est donc de topologie  $(1, 2)$ . Et enfin s'ils sont reliés par une liaison simple, il est de topologie  $(1, 3)$ .

Cependant notre modèle ne différencie pas les types de liaisons. Afin de pouvoir déterminer la topologie du sommet  $v$ , on utilise le nombre de doublets de son voisin pour en déduire le nombre de doublets du sommet initial. Le nombre de doublets d'un sommet, n'étant ni un hydrogène ni un halogène et ne possédant qu'un seul voisin est le même que celui de son voisin.

En effet, si son voisin a 4 doublets c'est qu'il ne possède aucune liaison multiple, par conséquent la liaison qui le relie au sommet  $v$  est une liaison simple. Puisqu'un seul électron est utilisé pour la liaison, il reste 3 doublets non-liants à  $v$ . De la même manière, si le voisin possède 3 doublets, c'est qu'une de ces liaisons est double. Dans ce cas, si le sommet  $v$  n'est ni un hydrogène ni un halogène, la liaison double est située entre  $v$  et son voisin. Le sommet  $v$  possède 2 doublets non-liants et sa topologie est de type  $(2, 1)$ . Enfin il en est de même si le voisin a 2 doublets. On est alors en présence d'une liaison triple et le sommet  $v$  est de topologie  $(1, 1)$ .

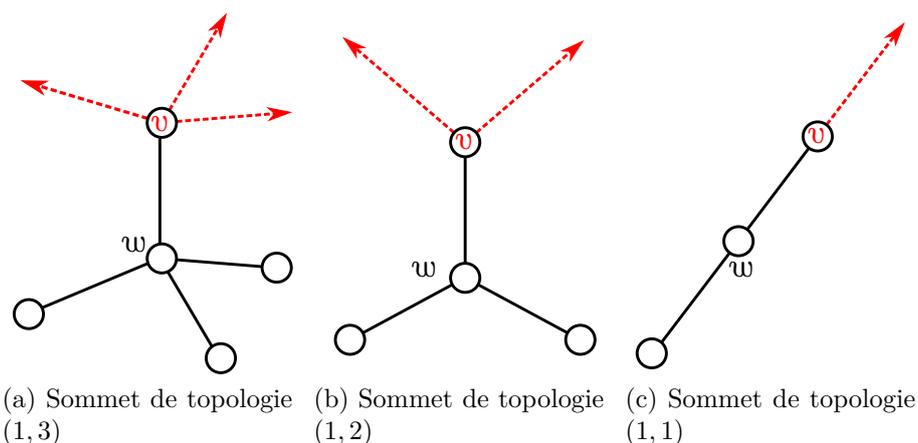


FIGURE 2.9 – Topologie des sommets ne possédant qu'un seul voisin

## 2.3 Algorithme général pour la recherche de la topologie d'un sommet

L'algorithme 4 prend en considération tous les cas énoncés précédemment afin de trouver la topologie de chaque sommet. Pour chaque sommet  $v$  du graphe  $G$ , on commence par créer le sous-graphe  $G_v$  constitué du sommet  $v$  et de ses voisins (2.1.2). On différencie ensuite le cas des sommets ayant un voisin unique. Pour chacun de ces sommets, on vérifie s'il n'est ni un hydrogène (2.2.1) ni un Halogène (2.2.2). S'il n'est aucun des deux, on récupère le nombre de doublets de son voisin pour lui donner le même (2.2.3). Dans le cas où les sommets ont plus d'un voisin, on calcule la moyenne des angles séparant ses voisins (2.1.2). Si la moyenne est d'environ  $109.28^\circ$ , soit il appartient à un cycle et dans ce cas son nombre de doublets est le même que ceux de ses

voisins (2.1.4), soit il n'appartient pas à un cycle et dans ce cas il possède 4 doublets. Si la moyenne des angles est d'environ  $120^\circ$ , il aura 3 doublets. Dans les autres cas, c'est-à-dire si la moyenne approche des  $180^\circ$ , il a 2 doublets.

---

**Algorithme 4 : Recherche Topologie**


---

**Entrées :** Graphe  $G$ , Sommet  $v$

**Sorties :** Nombre de doublets

```

1 si  $v$  n'a pas encore de topologie alors
2   Initialisation de  $G_v$ 
3    $n_v = \|E_v\|$ 
4   si  $n_v = 1$  alors
5     si  $symbol_v = H$  alors
6       |  $m_v = 1$ 
7     sinon si  $symbol_v = Cl$  ou  $symbol_v = F$  ou  $symbol_v = Br$  ou
8       |  $symbol_v = I$  alors
9         |  $m_v = 3$ 
10    sinon
11     |  $m_v = Topo(G, n) - n_v$ 
12    fin
13  sinon
14     $\bar{x}_v = \frac{2 * \sum \alpha_{(a,b)}}{n_v(n_v-1)}$ 
15    si  $\bar{x}_v \leq 109.28 + \epsilon$  alors
16      | si  $cycle_v = VRAI$  et  $(Topo(G, n_1) = 3$  ou  $Topo(G, n_2) = 3$  alors
17        |  $m_v = 3 - n_v$ 
18      | sinon
19        |  $m_v = 4 - n_v$ 
20      fin
21    sinon si  $\bar{x}_v \leq 120 + \epsilon$  alors
22      |  $m_v = 3 - n_v$ 
23    sinon
24      |  $m_v = 0$ 
25    fin
26 fin
27 retourner  $n_v + m_v$ 

```

---

Cet algorithme nous permet de déterminer la topologie de chaque sommet du graphe  $G$  en utilisant ses propriétés chimiques ou géométriques.

### 3 Expansion du substrat

La seconde étape de la construction de l'enveloppe est l'**expansion** du substrat, c'est-à-dire qu'on définit l'ensemble des sommets de l'enveloppe. Pour rappel, les sommets de l'enveloppe  $S$  doivent répondre à deux critères. Ils doivent être situés dans l'une des **directions libres** de l'un des sommets de  $G$  et à une distance équivalente à celle d'une liaison hydrogène de ce même sommet (environ  $1.8pm$ ), c'est-à-dire sur l'une de ses **positions libres**. En résumé, chaque sommet de  $S$  doit être situé à une position libre de l'un des sommets de  $G$ . L'ensemble des positions libres d'un sommet  $v$  de  $G$  est l'extension de  $v$  et est noté  $Ex_v$ . L'ensemble des extensions des sommets

de  $G$  est l'expansion de  $G$ , c'est-à-dire  $V_S$ .

À partir de la topologie d'un sommet, on connaît sa géométrie (tétraédrique, triangulaire ou linéaire). Et en fonction de la géométrie d'un sommet, on peut en déduire ses directions libres et ainsi ses positions libres. Une position libre d'un sommet  $v$  de  $G$  est l'ensemble des coordonnées spatiales résultant d'une translation appliquée sur  $v$  dans l'une de ses directions libres à une distance de  $1.8pm$ .

**Definition 23.** On note  $t_v(dis, dir)$ , la translation appliquée à un sommet  $v$  avec une distance  $dis$  et un vecteur directeur  $dir$ .

Cependant, dans le cas, il existe une infinité de solution pour ses directions libres. C'est pourquoi à chaque sommet  $v$ , on associe un vecteur dans l'espace 3D qu'on note  $normal_v$ . Ce vecteur est utilisé comme indicateur pour déterminer les directions libres les plus appropriées pour un sommet donné par rapport au reste du graphe.

Dans cette section, nous expliquons comment déterminer le vecteur  $normal_v$  de chaque sommet  $v$  de  $G$ . Puis, nous expliquons comment déterminer les directions libres des sommets de  $G$  en fonction de leur topologie afin de trouver les positions libres et ainsi de construire  $V_S$ .

### 3.1 Normale d'un sommet

À chaque sommet  $v$  du graphe  $G$ , on associe un vecteur directeur  $normal_v$ . Ce vecteur est une indication pour déterminer les directions libres d'un sommet en fonction de sa géométrie, et de sa topologie.

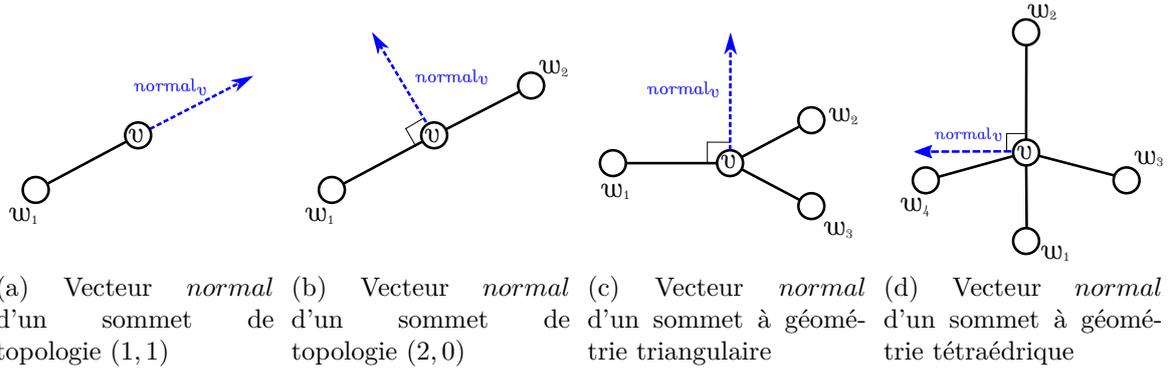
Pour rappel, un vecteur est défini par deux composantes :

- sa direction : celle de la droite qui porte le vecteur.
- son sens : oriente le vecteur.

Comme le montre la Figure 2.10, on distingue quatre cas. Le premier est le cas des sommets de topologie  $(1, 1)$ . Le vecteur  $normal_v$  d'un sommet de topologie  $(1, 1)$  a la même direction que celle reliant le sommet  $v$  et son voisin mais dans le sens opposé au voisin, c'est-à-dire dans la direction du doublet non-liant.

Le second est celui des sommets de topologie  $(2, 0)$ , le vecteur  $normal_v$  est l'un des vecteurs appartenant au plan perpendiculaire à la droite formée par les voisins du sommet. Plusieurs solutions sont possibles, on choisit le même vecteur  $normal$  que l'un des voisins du sommet. Dans le cas des sommets de géométrie triangulaire,  $normal_v$  est le vecteur de la normale du plan engendré par  $v$  et ses voisins. Pour finir, le vecteur  $normal_v$  quand  $v$  a une géométrie tétraédrique est la normale du plan engendré par  $v$  et deux de ses voisins.

Comme pour les sommets de topologie  $(2, 0)$ , il existe plusieurs solutions pour les sommets de topologie  $(1, 2)$  et  $(1, 3)$  puisqu'on ne possède que deux points du plan ( $v$  et son voisin). Ici aussi on choisit le vecteur  $normal_v$  d'un sommet  $v$  de topologie  $(1, 2)$  ou  $(1, 3)$  comme étant le même que son voisin.

FIGURE 2.10 – Position du vecteur  $normal_v$  en fonction des propriétés des sommets.

En reprenant les propriétés énoncées, l'algorithme 5 permet de calculer le vecteur  $normal_v$  pour un sommet  $v$  de  $G$ . Le sommet *pere* est utilisé pour les sommets de topologie (2,2) afin d'éviter de boucler indéfiniment entre deux sommets. Dans cet algorithme, la fonction  $NormalPlan(A, B, C)$  calcule la normale du plan engendré par les points A, B et C. On note  $w_v(i)$  comme le  $i^{eme}$  voisin de  $v$ .

---

**Algorithme 5 : RechercheNormale**


---

**Entrées :** Graphe  $G$ , Sommet  $v$ , Sommet *pere*

**Sorties :** Vecteur directeur  $Normal_v$  du sommet  $v$

```

1 si  $n_v = 1$  alors
2   | si  $m_v = 1$  alors
3   |   | retourner  $\overrightarrow{-vw_v(1)}$ 
4   | sinon
5   |   | retourner RechercheNormale( $G, w_v(1), v$ )
6   | fin
7 fin
8 si  $n_v + m_v = 2$  alors
9   | si  $w(1)_v \neq pere$  alors
10  |   | retourner RechercheNormale( $G, w_v(1), v$ )
11  | sinon
12  |   | retourner RechercheNormale( $G, w_v(2), v$ )
13  | fin
14 fin
15 si  $n_v + m_v = 3$  alors
16 |   | retourner NormalePlan( $v, w_v(1), w_v(2)$ )
17 fin
18 si  $n_v + m_v = 4$  alors
19 |   | retourner NormalePlan( $v, w_v(1), w_v(2)$ )
20 fin

```

---

### 3.2 Extension des sommets à géométrie tétraédrique

Les sommets possédant quatre doublets, c'est-à-dire les sommets de topologie (4,0), (3,1), (2,2) ou (1,3) ont une géométrie tétraédrique. On appellera  $\vec{d}_1$ ,  $\vec{d}_2$ ,  $\vec{d}_3$  et  $\vec{d}_4$  les quatre vecteurs indiquant la direction et le sens des doublets. Les directions libres de

ces sommets sont les doublets non-liants.

**Les sommets de topologie (4, 0)** n'ont pas de direction libre puisqu'ils ne possèdent pas de doublets non-liants. Leur extension est l'ensemble vide (voir Figure 2.11a).

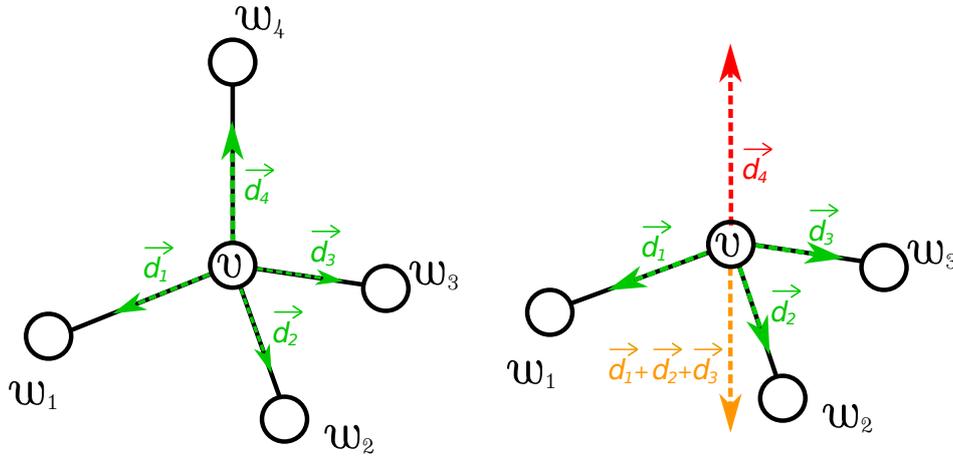
$\forall v \in V_G, topo_v = (4, 0)$  :

- $|E_v| = 0$
- $E_v = \{\emptyset\}$

**Les sommets de topologie (3, 1)** ont une seule direction libre, car ils n'ont qu'un seul doublet non-liant.

$\forall v \in V_G, topo_v = (3, 1)$  :

- $|E_v| = 1$
  - $E_v = \{t_v(1.8, \vec{d}_4)\}$
- avec  $\begin{cases} \vec{d}_1 = \frac{vw_v(1)}{\|vw_v(1)\|} \\ \vec{d}_2 = \frac{vw_v(2)}{\|vw_v(2)\|} \\ \vec{d}_3 = \frac{vw_v(3)}{\|vw_v(3)\|} \\ \vec{d}_4 = -(\vec{d}_1 + \vec{d}_2 + \vec{d}_3) \end{cases}$



(a) Directions libres d'un sommet (4, 0). (b) Directions libres d'un sommet (3, 1).

FIGURE 2.11 – Directions libres des sommets de topologie (4, 0) et (3, 1).

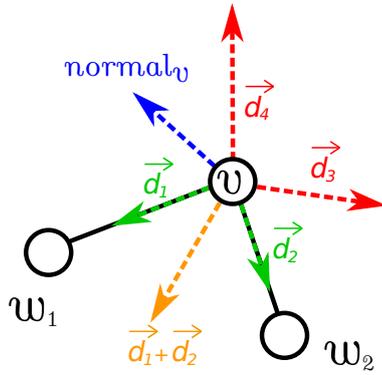
**Un sommet de topologie (2, 2)** a deux directions occupées par ses deux voisins dans  $G$  et il lui reste donc deux directions libres. La Figure 2.12 montre les deux directions libres ( $\vec{d}_3$  et  $\vec{d}_4$ ) qui sont situées dans un plan parallèle, noté  $P_2$ , à celui formé par les directions déjà utilisées ( $\vec{d}_1$  et  $\vec{d}_2$ ), noté  $P_1$ . De plus,  $P_2$  doit être situé à équidistance de  $\vec{d}_1$  et de  $\vec{d}_2$ .

On cherche la normale du plan  $P_2$  afin de déterminer  $\vec{d}_3$  et  $\vec{d}_4$ . Par définition,  $normal_v$  est la normale du plan  $P_1$ , on a donc  $normal_v \perp P_1$  d'où  $normal_v \in P_2$ . Comme on peut le voir sur la Figure 2.12a, on calcule  $d' \in P_2$  avec  $d' = \vec{d}_1 + \vec{d}_2$ . En effet puisque  $\|\vec{d}_1\| = 1$  et  $\|\vec{d}_2\| = 1$ ,  $d'$  est situé à équidistance de  $\vec{d}_1$  et  $\vec{d}_2$ . On

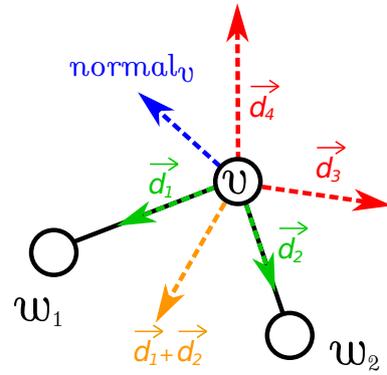
a alors  $normal_{P_2} = NormalePlan(\overrightarrow{normal_v}, \vec{d}')$ . Enfin, on applique une rotation d'axe  $normal_{P_2}$  sur  $\vec{d}'$  et d'angle  $\beta = \frac{360-109.28}{2} = 125.36$  (voir Figure 2.12b).

$\forall v \in V_G, topo_v = (2, 2) :$

$$\begin{aligned} & \bullet |E_v| = 2 \\ & \bullet E_v = \{t_v(1.8, \vec{d}_3), t_v(1.8, \vec{d}_4)\} \\ \text{avec } \begin{cases} \vec{d}_1 &= \frac{\overrightarrow{vw_v}(1)}{\|\overrightarrow{vw_v}(1)\|} \\ \vec{d}_2 &= \frac{\overrightarrow{vw_v}(2)}{\|\overrightarrow{vw_v}(2)\|} \\ \vec{d}_3 &= rotation(normal_{P_2}, d', \beta) \\ \vec{d}_4 &= -(\vec{d}_1 + \vec{d}_2 + \vec{d}_3) \end{cases} \end{aligned}$$



(a) Normale du plan  $P_2$ .



(b) Directions libres d'un sommet  $(2, 2)$ .

FIGURE 2.12 – Directions libres d'un sommet de topologie  $(2, 2)$ .

[Trouver comme faire des plans.](#)

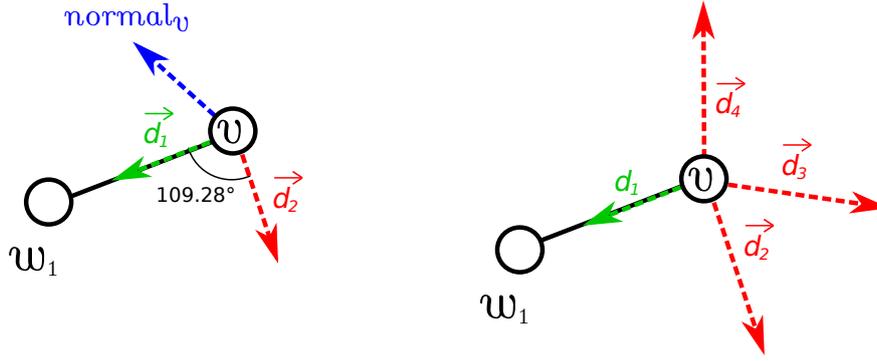
**Un sommet de topologie  $(1, 3)$**  ne possède qu'un seul voisin et a donc trois directions libres. Cependant, comme expliqué dans la section 3.1, il y a une infinité de solutions. C'est pourquoi on utilise  $normal_v$  comme axe de rotation pour trouver la direction de la première direction libre, comme sur la Figure 2.13. Ensuite, on utilise simplement les mêmes méthodes que pour les sommets de topologies  $(2, 2)$  puis  $(3, 1)$ .

$\forall v \in V_G, topo_v = (1, 3) :$

$$\begin{aligned} & \bullet |E_v| = 3 \\ & \bullet E_v = \{t_v(1.8, \vec{d}_2), t_v(1.8, \vec{d}_3), t_v(1.8, \vec{d}_4)\} \\ \text{avec } \begin{cases} \vec{d}_1 &= \frac{\overrightarrow{vw_v}(1)}{\|\overrightarrow{vw_v}(1)\|} \\ \vec{d}_2 &= rotation(normal_v, d_1, 109.28) \\ \vec{d}_3 &= rotation(normal_{P_2}, d', \beta) \\ \vec{d}_4 &= -(\vec{d}_1 + \vec{d}_2 + \vec{d}_3) \end{cases} \end{aligned}$$

### 3.3 Extension des sommets à géométrie triangulaire

Les sommets possédant trois doublets, c'est-à-dire les sommets de topologie  $(3, 0)$ ,  $(2, 1)$  et  $(1, 2)$  ont une géométrie triangulaire. On appellera  $\vec{d}_1$ ,  $\vec{d}_2$  et  $\vec{d}_3$  les trois



(a) Première direction libre d'un sommet de topologie (1,3) (b) Directions libres d'un sommet de topologie (1,3).

FIGURE 2.13 – Directions libres d'un sommet de topologie (1,3).

vecteurs indiquant la direction et le sens des doublets. Comme pour les sommets à géométrie tétraédrique, les directions libres des sommets à géométrie triangulaire sont les doublets non-liants. Cependant, étant donné que tous ces vecteurs sont situés dans le même plan, d'autres interactions sont possibles dans les directions perpendiculaires au plan. Leur géométrie est comparable au sommet possédant cinq doublets. On appellera  $\vec{d}_4$  et  $\vec{d}_5$  les deux directions libres perpendiculaires au plan formé par le sommet  $v$  et ses voisins. Ces deux directions libres sont toujours dans la même direction que le vecteur  $normal_v$  du sommet.

**Un sommet  $v$  de topologie (3,0)** n'a pas de doublet non liant. Il a seulement deux directions libres perpendiculaires au plan engendré par ses voisins, comme le montre la Figure 2.14a.

**Cas 5 :**  $\forall v \in V_G, topo_v = (3,0)$  :

- $|E_v| = 2$
  - $E_v = \{t_v(1.8, \vec{d}_4), t_v(1.8, \vec{d}_5)\}$
- avec  $\begin{cases} \vec{d}_4 = normal_v \\ \vec{d}_5 = -normal_v \end{cases}$

**Un sommet  $v$  de topologie (2,1)** a un doublet non-liant, si on ajoute les deux directions libres perpendiculaires, il a donc trois directions libres au total.

$\forall v \in V_G, topo_v = (2,1)$  :

- $|E_v| = 3$
  - $E_v = \{t_v(1.8, \vec{d}_3), t_v(1.8, \vec{d}_4), t_v(1.8, \vec{d}_5)\}$
- avec  $\begin{cases} \vec{d}_1 = \frac{vw_v(1)}{\|vw_v(1)\|} \\ \vec{d}_2 = \frac{vw_v(2)}{\|vw_v(2)\|} \\ \vec{d}_3 = -(\vec{d}_1 + \vec{d}_2) \\ \vec{d}_4 = normal_v \\ \vec{d}_5 = -normal_v \end{cases}$

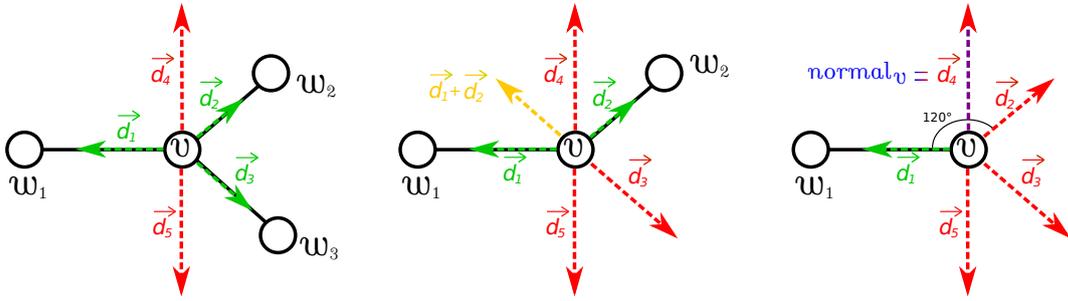
**Un sommet de topologie (1, 2)** qui ne possède qu'un voisin et a donc deux doublets non-liants. Avec les deux directions libres perpendiculaires, un sommet de topologie (1, 2) a un total de quatre directions libres. Cependant, comme pour les sommets de topologie (1, 3), il n'y a pas une infinité de solutions possibles pour les directions libres des doublets non-liants. Afin de trouver les directions libres les plus intéressantes, chimiquement parlant, on utilise le vecteur  $normal_v$  pour les déterminer. En effet, puisque  $normal_v$  d'un sommet de topologie (1, 2) est la même que celle de son voisin, les deux directions libres sont ainsi dans le même plan que les sommets de  $G$  situés à une distance inférieure ou égale à 2 de  $v$ . On effectue une rotation de  $120^\circ$  du voisin de  $v$  en utilisant  $normal_v$  comme axe de rotation (voir Figure 2.14c).

$\forall v \in V_G, topo_v = (1, 2) :$

- $|E_v| = 4$
- $E_v = \{t_v(1.8, \vec{d}_2), t_v(1.8, \vec{d}_3), t_v(1.8, \vec{d}_4), t_v(1.8, \vec{d}_5)\}$

avec

$$\begin{cases} \vec{d}_1 = \frac{\vec{vw}_v(1)}{\|\vec{vw}_v(1)\|} \\ \vec{d}_2 = rotation(normal_v, d_1, 120) \\ \vec{d}_3 = rotation(normal_v, d_1, -120) \\ \vec{d}_4 = normal_v \\ \vec{d}_5 = -normal_v \end{cases}$$



(a) Directions libres d'un sommet (3, 0).

(b) Directions libres d'un sommet (2, 1).

(c) Directions libres d'un sommet (1, 2)

FIGURE 2.14 – Directions libres des sommets à géométrie triangulaire.

### 3.4 Extension des sommets à géométrie linéaire

Il existe deux types de sommets à géométrie linéaire. Les sommets de topologie (2, 0) et les sommets de topologie (1, 1). On appellera  $\vec{d}_1$  et  $\vec{d}_2$  les deux vecteurs indiquant la direction et le sens des doublets.

**Un sommet de topologie (2, 0)** n'a pas de doublets non-liants. Cependant dans la zone où la force exercée par ses électrons est la plus faible, c'est-à-dire dans le plan perpendiculaire à la droite formée par ses voisins passant par  $v$ , d'autres interactions sont possibles. C'est pourquoi pour chaque sommet  $v$  de topologie

$(2, 0)$  on définit quatre directions libres autour du sommet qu'on appellera  $\vec{d}_3$ ,  $\vec{d}_4$ ,  $\vec{d}_5$  et  $\vec{d}_6$ . Pour les trouver, on utilise la normale du sommet  $v$ ,  $normal_v$ , pour déterminer la première direction libre, ainsi que  $\vec{d}_1$  comme axe de rotation de la manière suivante.

$\forall v \in V_G$ ,  $topo_v = (2, 0)$  :

- $|E_v| = 4$
- $E_v = \{t_v(1.8, \vec{d}_3), t_v(1.8, \vec{d}_4), t_v(1.8, \vec{d}_5), t_v(1.8, \vec{d}_6)\}$

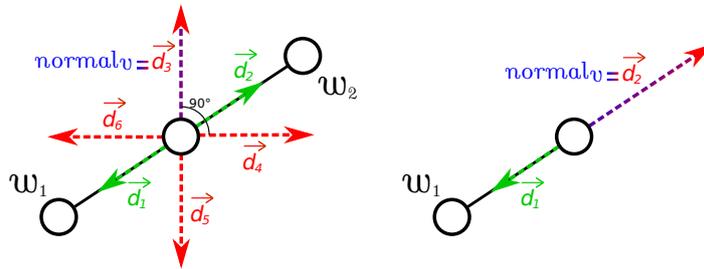
avec  $\left\{ \begin{array}{l} \vec{d}_1 = \frac{\overrightarrow{vw}_v(1)}{\|\overrightarrow{vw}_v(1)\|} \\ \vec{d}_2 = \frac{\overrightarrow{vw}_v(2)}{\|\overrightarrow{vw}_v(2)\|} \\ \vec{d}_3 = normal_v \\ \vec{d}_4 = rotation(\vec{d}_1, \vec{d}_3, 90) \\ \vec{d}_5 = -normal_v \\ \vec{d}_6 = -\vec{d}_4 \end{array} \right.$

**Un sommet  $v$  de topologie  $(1, 1)$**  a une seule direction libre, qui est dans la même direction et le même sens que son vecteur  $normal_v$ .

$\forall v \in V_G$ ,  $topo_v = (1, 1)$  :

- $|E_v| = 1$
- $E_v = \{t_v(1.8, \vec{d}_2)\}$

avec  $\vec{d}_2 = normal_v$ .



(a) Directions libres d'un sommet  $(2, 0)$ .

(b) Directions libres d'un sommet  $(1, 1)$

FIGURE 2.15 – Directions libres des sommets à géométrie linéaire.

### 3.5 Récapitulatif

Pour chaque sommet du substrat  $G$ , on a calculé son extension, c'est-à-dire l'ensemble des positions libres qu'il engendre. Chacune de ces positions est un sommet de l'enveloppe  $S$  qui aura pour coordonnées dans l'espace les coordonnées de la position libre.  $V_S$  est l'ensemble des positions libres des sommet de  $G$ .

Dans le cas des sommets à géométrie tétraédrique, c'est-à-dire les sommets de topologie  $(4, 0)$ ,  $(3, 1)$ ,  $(2, 2)$  ou  $(1, 3)$ , chacun des doublets est associé à un vecteur allant

du sommet (au centre du tétraèdre) vers les sommets du polyèdre dans lequel il est inscrit. Les vecteurs des doublets liants servent de départ pour calculer les autres vecteurs et ainsi les directions libres. Dans le cas des sommets de topologie  $(1, 3)$ , comme un seul des vecteurs est déterminé puisque les sommets ne possèdent qu'un seul voisin, il existe une infinité de solutions pour les vecteurs restants. Afin de pouvoir déterminer la solution la plus appropriée, le vecteur *normal* du sommet est également utilisé.

Dans le cas des sommets à géométrie triangulaire, c'est-à-dire les sommets de topologie  $(3, 0)$ ,  $(2, 1)$  ou  $(1, 2)$ , chacun des doublets est associé à un vecteur allant du sommet (au centre du triangle) vers les sommets du polygone dans lequel il est inscrit. Ici encore, les vecteurs associés aux doublets liants servent de départ pour calculer les autres vecteurs. Pour les sommets de topologie  $(1, 2)$ , le vecteur *normal* du sommet est aussi utilisé. De plus, chacun de ces sommets a également des directions libres dans les deux sens de la droite perpendiculaire au triangle dans lequel lui et ses voisins sont inscrits.

Enfin dans le cas des sommets à géométrie linéaire, c'est-à-dire les sommets de topologie  $(2, 0)$  ou  $(1, 1)$ , chaque doublet est associé à un vecteur allant dans les deux sens de la droite passant par lui et ses voisins. Les sommets de topologie  $(1, 1)$  ont donc une direction libre dans le sens opposé à celui du voisin du sommet. Malgré le fait que les deux vecteurs des sommets de topologie  $(2, 0)$  soient déjà associés aux voisins du sommet, ils possèdent quatre directions libres. Ces directions libres sont les vecteurs associés au carré perpendiculaire à la droite dans laquelle lui et ses voisins sont inscrits. Le vecteur *normal* du sommet est la première direction libre et est ensuite utilisé pour déterminer les vecteurs allant vers les autres sommets du carré.

## 4 Construction des arêtes de l'enveloppe

L'expansion du substrat  $G$  a permis de construire l'ensemble des sommets de l'enveloppe. La prochaine étape est de construire les arêtes de  $S$  de telle sorte que les propriétés énoncées dans la section 1.2 soient respectées.

### 4.1 Propriété attendue

On veut que  $S$  englobe géométriquement le substrat  $G$ . Aucune arête de  $S$  ne doit traverser la structure définie par  $G$ . Les sommets de  $S$  ont été positionnés dans les directions libres des sommets de  $G$ . Dans la plupart des cas, ces directions libres partent dans le sens opposé au substrat  $G$ . Ainsi les sommets de  $S$  sont, pour la plupart, situés autour de la structure définie par  $G$ . En construisant une enveloppe autour des sommets de  $S$ , aucune des arêtes de  $S$  ne traversera  $G$ .

On définit l'*enveloppe* d'un nuage de points comme un ensemble d'arêtes définissant une structure dans laquelle tous les points sont inclus. Il existe deux types d'enveloppes, les enveloppes convexes et les enveloppes concaves. Une *enveloppe convexe*, comme nous le montre la Figure 2.16a, est l'enveloppe qui minimise la zone qui contient tous les points, sans que l'angle entre deux arêtes voisines ne dépasse  $180^\circ$ . En conséquence si on prend deux points quelconque du nuage, le segment qui les joint est contenu dans l'objet défini par l'enveloppe. Pour un ensemble de points donné, l'enveloppe convexe est unique.

Une *enveloppe concave* est une enveloppe qui délimite une zone contenant tous les points, cependant, contrairement à l'enveloppe convexe, il n'y a pas de contrainte d'angle. Une enveloppe concave pour un ensemble de points donnés n'est pas unique. En effet, toutes les enveloppes, à l'exception de l'enveloppe convexe, sont considérées comme concaves. La Figure 2.16b nous montre deux exemples d'enveloppes concaves pour un même nuage de points.

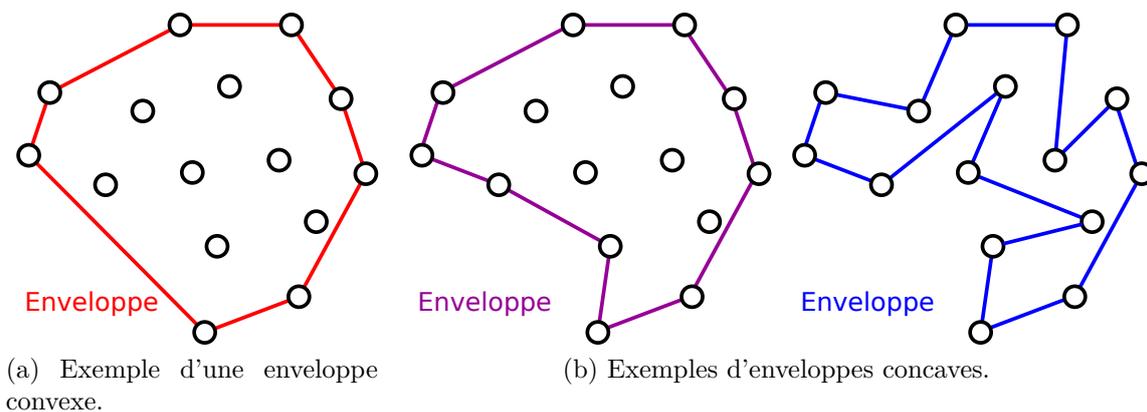


FIGURE 2.16 – Exemple d'enveloppe d'un nuage de points en 2D.

Puisque dans notre cas, on veut que l'enveloppe garde une forme assez proche de celle du substrat, l'enveloppe la plus adaptée est une enveloppe concave. Cependant, on veut malgré tout que tous les sommets du substrat soient à l'intérieur de l'enveloppe. Les sommets de l'expansion trop proches du substrat ne seront pas insérés dans l'enveloppe.

## 4.2 Alpha shape

La méthode *Alpha Shape* (Edelsbrunner et al. (1983); Edelsbrunner and E.P.Mücke (1994)) est une méthode qui permet de construire des enveloppes autour d'un nuage de points. Elle peut être utilisée aussi bien pour générer des enveloppes concaves que l'enveloppe convexe. Cette méthode s'effectue en deux étapes. La première étape est une triangulation effectuée sur l'ensemble des points à traiter. La deuxième consiste à retirer certains triangles afin d'ajuster l'enveloppe concave souhaitée.

### 4.2.1 Delaunay

La triangulation utilisée dans la méthode Alpha-shape est une triangulation de Delaunay (Delaunay (1924)). Cette triangulation permet d'éviter les triangles « allongés » en maximisant les angles les plus petits. On réalise dans un premier temps une triangulation quelconque sur l'ensemble des points de départ puis on vérifie si tous les triangles respectent la condition de Delaunay.

Un triangle est dit de Delaunay si aucun point n'est à l'intérieur du cercle circonscrit au triangle. Pour que tous les triangles respectent cette condition, on peut utiliser différentes techniques. L'une d'entre elles est la méthode de basculement. Pour chaque

arête de la triangulation appartenant à deux triangles, on va vérifier si ces deux triangles respectent la condition de Delaunay. Si ce n'est pas le cas, on bascule l'arête, c'est-à-dire qu'on supprime l'arête pour en reconstruire une autre entre les deux autres sommets des triangles. La Figure 2.17 est un exemple de basculement. On cherche à déterminer si l'arête rouge de la Figure 2.17a doit être conservée ou non. On voit sur la Figure 2.17b qu'au moins un des triangles auxquels elle appartient ne respecte pas la condition de Delaunay puisque le dernier sommet du second triangle est inclus dans le cercle circonscrit au premier. On supprime donc cette arête pour la remplacer par l'arête de l'autre diagonale, comme on le voit sur la Figure 2.17c. Les deux nouveaux triangles respectent la condition de Delaunay et sont conservés dans la triangulation.

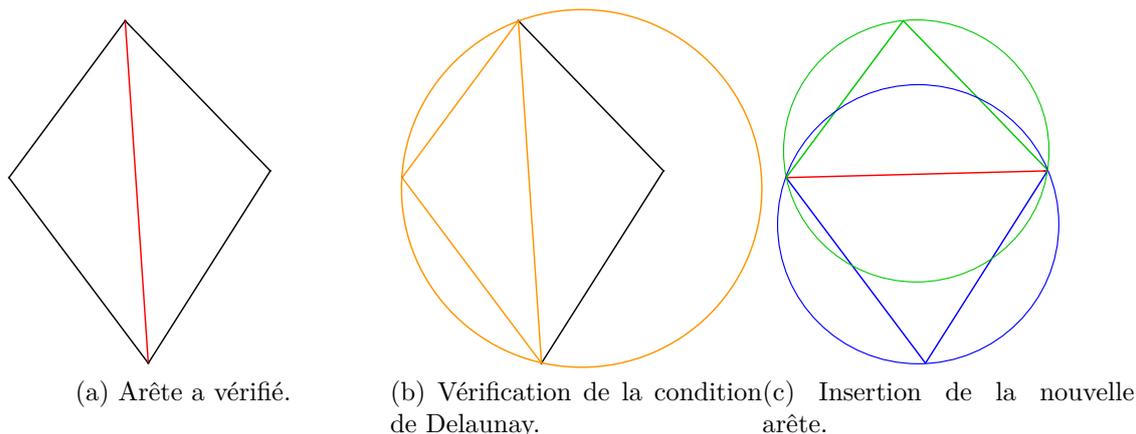


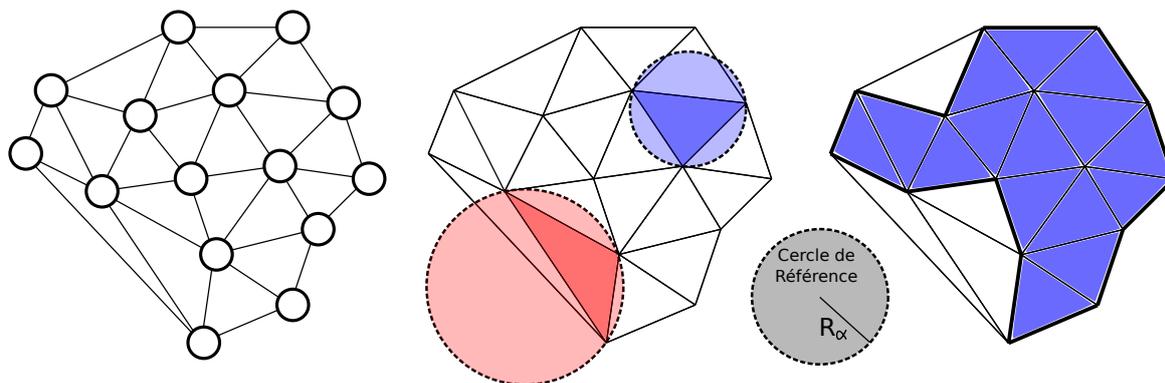
FIGURE 2.17 – Exemple d'une étape de la triangulation de Delaunay.

La triangulation de Delaunay est applicable aux dimensions supérieures. Dans le cas d'un nuage de points dans un espace en 3D, on ne s'intéresse plus aux triangles mais aux tétraèdres. Pour qu'un tétraèdre soit de Delaunay, aucun point ne doit être à l'intérieur de la sphère circonscrite au tétraèdre formé par quatre points de la triangulation.

#### 4.2.2 Alpha-shape

Si on conserve exclusivement les arêtes extérieures, en 2D, ou les triangles extérieurs, en 3D, de l'espace triangulé, on obtient l'enveloppe convexe du nuage de points. La méthode *Alpha-shape* permet d'affiner l'enveloppe en retirant certaines arêtes ou certains triangles de la triangulation en fonction d'un paramètre  $\alpha$ . L'enveloppe devient une enveloppe concave du nuage de points. Le paramètre  $\alpha$  est comparé au rayon des cercles circonscrits aux triangles de la triangulation. Comme on peut le voir sur la Figure 2.18, le paramètre  $\alpha$  est utilisé comme référence. Si la taille du cercle est supérieure à la taille du cercle de référence, le triangle est retiré de la triangulation. À l'inverse si la taille du cercle est inférieure à celle de référence, le triangle est conservé dans la forme alpha. En 3D, le paramètre  $\alpha$  est utilisé comme rayon d'une sphère de référence qui est comparée aux sphères circonscrites des tétraèdres.

Ainsi, si le paramètre  $\alpha$  tend vers l'infini, la forme alpha sera l'enveloppe convexe. À l'inverse si le paramètre  $\alpha$  tend vers 0, aucun triangle ne sera conservé. C'est pourquoi le choix du paramètre  $\alpha$  est important. Dans notre cas, nous avons déterminé une fourchette pour le paramètre  $\alpha$  par apprentissage. Dans la mesure où la distance séparant les atomes est similaire d'une molécule à l'autre (puisqu'elle dépend du type

FIGURE 2.18 – Représentation de la méthode *Alpha-shape*.

des atomes eux-mêmes), les distances séparant deux points d'une expansion sont également proches d'une expansion à l'autre. C'est pourquoi on a pu déterminer une fourchette pour le paramètre  $\alpha$  (voir annexe B).

## 5 Conclusion

L'enveloppe est une structure spécifique au substrat. Bien qu'elle ne soit pas assimilable à une molécule, c'est une structure construite dans une logique moléculaire, ainsi chacun de ses sommets est situé dans une direction où les atomes du substrat peuvent encore interagir avec d'autres molécules.

La construction de cette enveloppe se fait en plusieurs étapes consécutives. La première consiste à déterminer la topologie des sommets qui composent le substrat, c'est-à-dire de déterminer le nombre de directions dans lesquelles chaque sommet peut encore créer des interactions avec d'autres molécules. La méthode utilisée pour définir la topologie d'un sommet est basée sur des règles qui dépendent du type d'atome que représente le sommet et de la position de ses voisins autour de lui. Dans la majorité des cas les règles établies permettent donc de déterminer la topologie d'un sommet.

L'étape suivante est de trouver les positions des sommets de l'enveloppe en utilisant la topologie des sommets. Tous ses sommets sont placés à une distance équivalente à celle d'une liaison hydrogène du substrat. Une amélioration possible serait de déterminer la distance en fonction du type de liaison auquel le sommet peut participer.

Enfin la dernière étape permet de déterminer les arêtes de l'enveloppe en construisant une enveloppe concave à partir des sommets trouvés. Le choix de l'enveloppe concave utilisée est laissé libre à l'utilisateur bien qu'une fourchette pour le paramètre ait été établie par apprentissage. Un travail intéressant serait d'affiner le choix du paramètre afin de déterminer la meilleure enveloppe pour le substrat.

L'intérêt de cette enveloppe est qu'elle sert de base pour construire des cages spécifiques pour le substrat donné.



# Chapitre 3

## Intégration des motifs liants

---

1	Intégration d'un motif moléculaire . . . . .	44
2	Construction des liaisons aromatiques . . . . .	51
3	Construction des Liaisons Hydrogènes . . . . .	54
4	Conclusion . . . . .	60

---

*Dans le chapitre précédent, nous avons montré comment construire une enveloppe à partir d'un substrat donné. Dans ce chapitre nous expliquons comment, à partir de cette enveloppe, nous intégrons des motifs moléculaires liants. Ces motifs feront partie des graphes moléculaires représentant les cages et serviront de points de liaisons entre le substrat et les cages. Les motifs moléculaires sont des sous-graphes moléculaires, l'ensemble des sommets qui les composent sont assimilables à des atomes dans un espace 3D et les arêtes représentent des liaisons covalentes. Les graphes moléculaires représentant les cages sont des assemblages de motifs moléculaires.*

*Les motifs que nous intégrons ici, sont assimilables à des ensembles d'atomes pouvant être impliqués dans des liaisons inter-moléculaires. Deux types de motifs vont nous intéresser ici : les motifs aromatiques et les motifs hydrogènes. Les premiers sont impliqués dans les liaisons aromatiques. Il s'agit de cycles ou de morceaux de cycles aromatiques. Le but est que les motifs aromatiques puissent interagir avec les cycles aromatiques du substrat. Les deuxièmes sont des ensembles d'atomes pouvant être impliqués dans des liaisons hydrogènes. Là encore, il est important de déterminer quels sont les ensembles d'atomes dans le substrat qui peuvent être impliqués dans des liaisons hydrogènes afin de définir l'emplacement de ces motifs dans l'enveloppe.*

*Dans un premier temps, nous expliquerons quelles sont les étapes générales de l'intégration d'un motif à l'enveloppe de manière à garder une cohérence au fil des intégrations. Ensuite, nous décrirons plus particulièrement l'intégration de chaque type de motifs (aromatiques puis hydrogènes). Pour chacun d'entre eux, nous expliquerons comment déterminer les zones de l'enveloppe où ils peuvent être intégrés, ainsi que la manière dont nous choisissons parmi les motifs d'un même type celui à intégrer.*

## 1 Intégration d'un motif moléculaire

Dans le chapitre 1, nous avons défini différents types de motifs moléculaires pouvant être intégrés dans l'enveloppe. Intégrer un motif consiste à remplacer un sous-graphe de l'enveloppe par un autre sous-graphe qui est le motif lui-même. Les motifs sont intégrés dans l'enveloppe de manière successive. Chaque sommet a une *priorité*. Les sommets de l'enveloppe  $S$  de départ, ont une priorité de 0 car ils ne remplissent pas les conditions d'un graphe moléculaire. Les sommets dits *saturés* (Définition 13) sont de priorité 2 car ils remplissent toutes les conditions d'un graphe moléculaire. Ce sont les sommets situés à l'intérieur des motifs moléculaires. Enfin les sommets de priorité 1 sont les sommets dits *insaturés* des motifs. Ce sont les sommets qui font la jonction entre les sommets de priorité 0 et les sommets de priorité 2.

Dans cette section, nous expliquons comment sont intégrés les motifs afin de garder une cohérence au fil des intégrations successives.

### 1.1 Étapes d'intégration d'un motif

Étant donné le graphe  $S = (V_S, E_S)$  de l'enveloppe et un motif  $M$ , l'intégration du motif s'effectue en quatre étapes :

1. Déterminer le sous-graphe  $G_1$  de  $S$  tel que  $G_1$  est le sous-graphe induit par les sommets de  $S$  à remplacer par le motif  $M$ .
2. Déterminer  $\Gamma(G_1)$  tel que  $\Gamma(G_1)$  est l'ensemble des sommets de  $S$  de priorité 0 ou 1 qui ont un voisin dans  $G_1$  mais n'appartiennent pas à  $G_1$ .
3. Suppression de  $G_1$  pour le remplacer par  $M$ .
4. Rattachement de  $M$  au reste de l'enveloppe. Pour chaque sommet  $v$  de  $\Gamma(G_1)$ , une arête est ajoutée entre  $v$  et le sommet  $m \in \text{bordure}(M)$  le plus proche de  $v$  géométriquement.

La *bordure* d'un motif est constituée de tous les sommets de priorité 1 auxquels peuvent être rattachés les sommets du reste de l'enveloppe.

Les Figures 3.1, 3.2 et 3.3 récapitulent les étapes de l'intégration d'un motif  $M$  dans une enveloppe  $S$ . Dans cet exemple, le motif à intégrer est un motif triangulaire comprenant quatre sommets dont le seul sommet saturé est le sommet situé au centre du motif. Ce sommet est donc de priorité 2 alors que les autres sommets du motif sont automatiquement de priorité 1.

La Figure 3.1a montre l'état de l'enveloppe  $S$  avant intégration du motif. La Figure 3.1b, quant à elle, montre l'endroit où le motif va être intégré par rapport à l'enveloppe. Dans cet exemple, certains sommets de l'enveloppe sont en « collision » avec les sommets du motif. Deux sommets sont dits en « collision » s'ils sont situés à une distance géométrique inférieure à la distance minimale pouvant exister entre deux sommets d'un graphe moléculaire (soit  $D_{min}$  définie dans le chapitre 1). Les sommets de  $G_1$  sont les sommets de l'enveloppe en collision avec ceux du motif.

Dans cet exemple, quatre sommets de l'enveloppe sont en collision avec des sommets du motif. Le graphe  $G_1$ , montré sur la Figure 3.2a, est donc composé de ses quatre sommets. L'étape suivante est de définir l'ensemble  $\Gamma(G_1)$  associé au sous-graphe  $G_1$ .

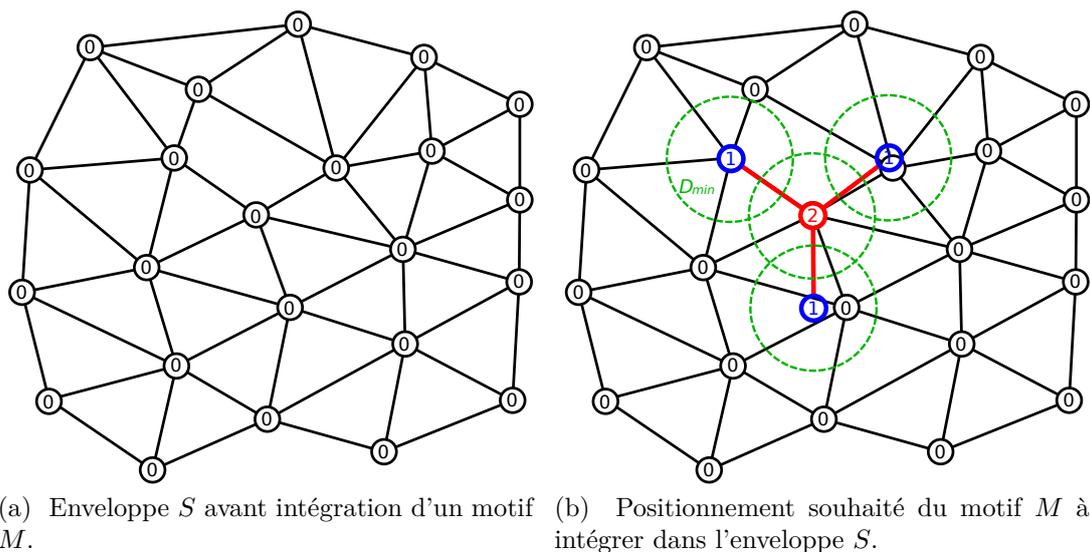


FIGURE 3.1 – Exemple de positionnement d'un motif dans une enveloppe.

Cet ensemble est représenté sur la Figure 3.2b. Chaque sommet appartenant à  $\Gamma(G_1)$  a une arête dans  $S$  qui le relie à l'un des sommets de  $G_1$ .

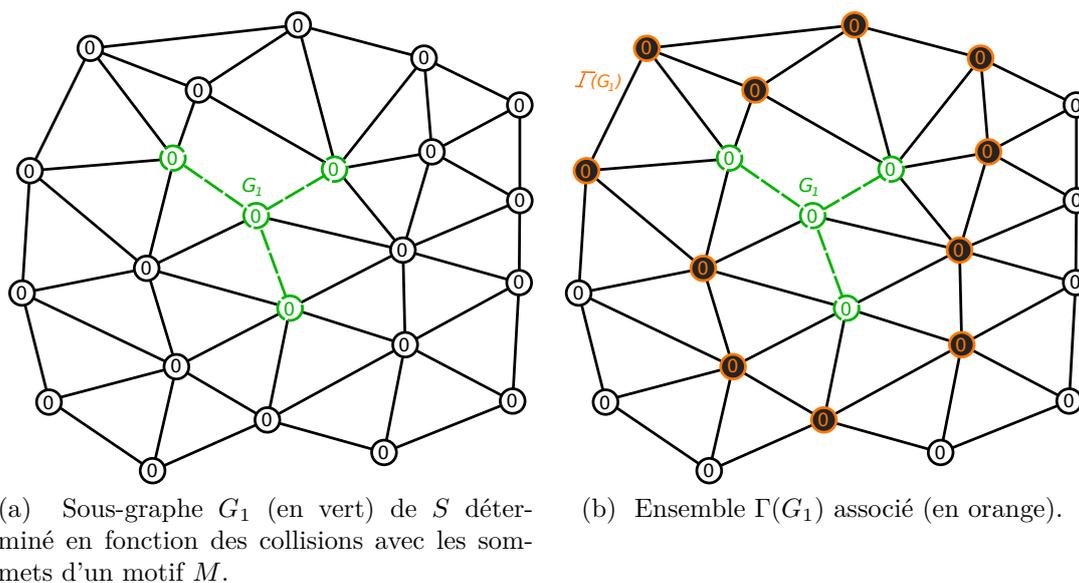


FIGURE 3.2 – Exemple des étapes 1 et 2 de l'intégration d'un motif.

Enfin, la Figure 3.3 regroupe les étapes trois et quatre de l'intégration du motif. La Figure 3.3a montre la suppression du sous-graphe  $G_1$  dans l'enveloppe  $S$ . Après suppression de  $G_1$ , le motif  $M$  peut être intégré dans l'enveloppe. C'est l'étape du rattachement illustrée dans la Figure 3.3b.

Dans l'exemple précédent, tous les sommets de l'enveloppe  $S$  sont des sommets de priorité 0. Puisque ces sommets ne respectent aucune des contraintes liées aux graphes moléculaires, ils n'appartiennent pas à la solution finale. Ils peuvent donc être supprimés et remplacés par les sommets du motif s'ils rentrent en collision.

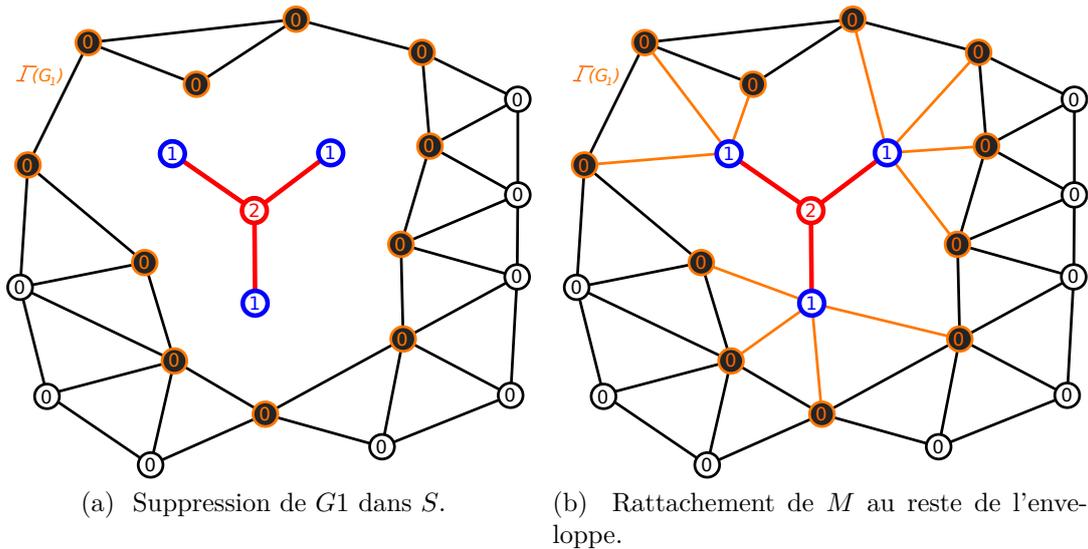


FIGURE 3.3 – Exemple des étapes 3 et 4 de l'intégration d'un motif.

Cependant la question se pose pour les sommets de priorité 1 et 2. En effet ces sommets sont une partie de la solution, ils ne peuvent donc pas être supprimés ou déplacés sans crainte de briser des contraintes. Les sommets saturés, ceux de priorité 2, ne peuvent pas être modifiés, c'est-à-dire que leur coordonnées, leurs arêtes et les coordonnées de leurs voisins doivent être conservés. Les sommets de priorité 1 font la jonction entre les sommets de priorité 2 et les sommets de priorité 0, ils ont donc au moins un voisin de priorité 2. Ils ne peuvent donc pas être supprimés ou déplacés et seules les arêtes qui les relient à d'autres sommets de priorité 0 ou 1 peuvent être modifiées.

Pour faire face à ce problème, la notion de recouvrement a été ajoutée. Un sommet de l'enveloppe est dit *recouvert* par un sommet d'un motif, si les deux sommets ont les mêmes coordonnées dans l'espace. Afin qu'un motif  $M$  puisse être intégré, il est nécessaire que tous les sommets de  $G_1$  de priorité 1 ou 2 soient recouverts par un sommet de  $M$ . Dans le cas contraire, le motif ne pourra pas être intégré dans l'enveloppe.

Dans la suite, nous continuons l'exemple précédent en intégrant un second motif, identique au premier mais à un emplacement différent.

La Figure 3.4 montre deux exemples de positionnement du motif. Dans les deux cas, certains sommets de  $G_1$  ont une priorité de 1 ou 2. Dans la Figure 3.4a, un des sommets de  $G_1$  de priorité 2 n'est pas recouvert par l'un des sommets de  $M$ . Le sommet saturé ne peut pas être supprimé ou modifié, mais il est malgré tout en collision avec un sommet du motif, les deux ne peuvent donc pas être conservés. En conséquent, le motif ne pourra pas être intégré. Dans la Figure 3.4b, tous les sommets de  $G_1$  de priorité 1 et 2 sont recouverts par un sommets de  $M$ , l'intégration est donc possible.

Si un sommet de  $G_1$  est un sommet de priorité 2, il est alors fusionné avec le sommet du motif qui le recouvre. Les arêtes qui le reliaient au reste de l'enveloppe sont conservées et l'arête qui était recouverte par celle du motif doit être rattachée au sommet du motif. Le sommet de priorité 2 est alors retiré de  $G_1$ .

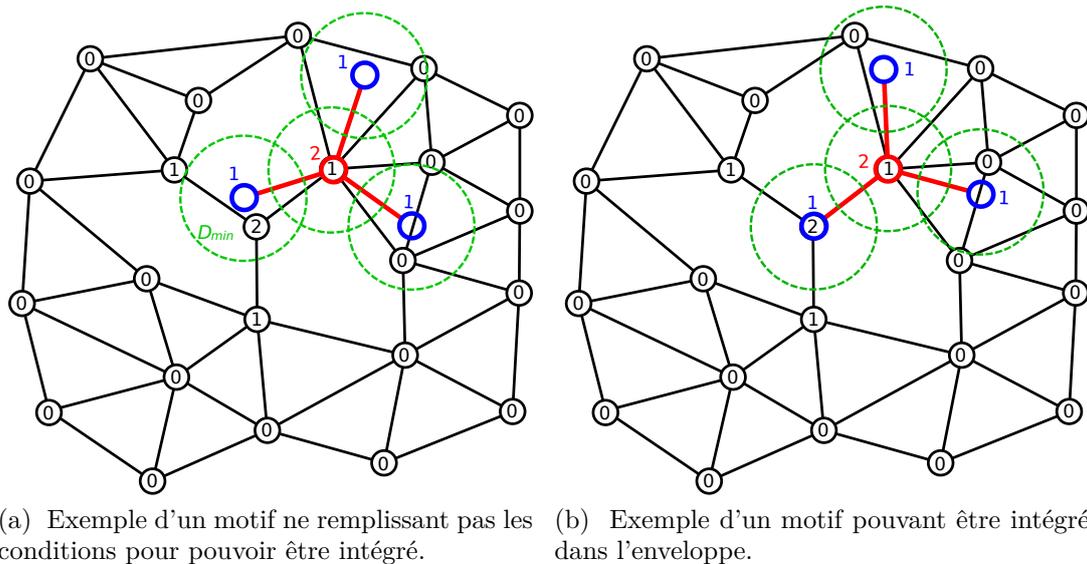


FIGURE 3.4 – Exemple de deux positionnements d'un même motif dans une enveloppe.

Si un sommet de  $G_1$  est un sommet de priorité 1, les arêtes qui le relient aux sommets de priorité 2 sont conservées en étant rajoutées au sommet du motif qui le recouvre.

La Figure 3.5a montre le sous-graphe  $G_1$  obtenu dans ces conditions. Le sommet de priorité 2 a été retiré de  $G_1$  car il a été fusionné avec les sommets du motif. La Figure 3.5b met en avant l'ensemble  $\Gamma(G_1)$  associé. Seuls les voisins de priorité 0 et 1 sont ajoutés à  $\gamma$ .

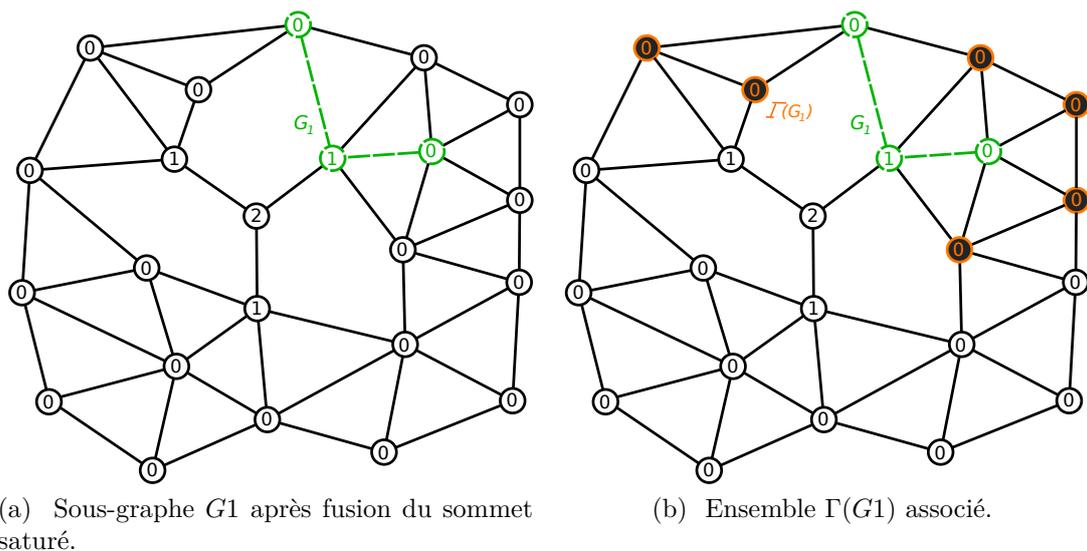
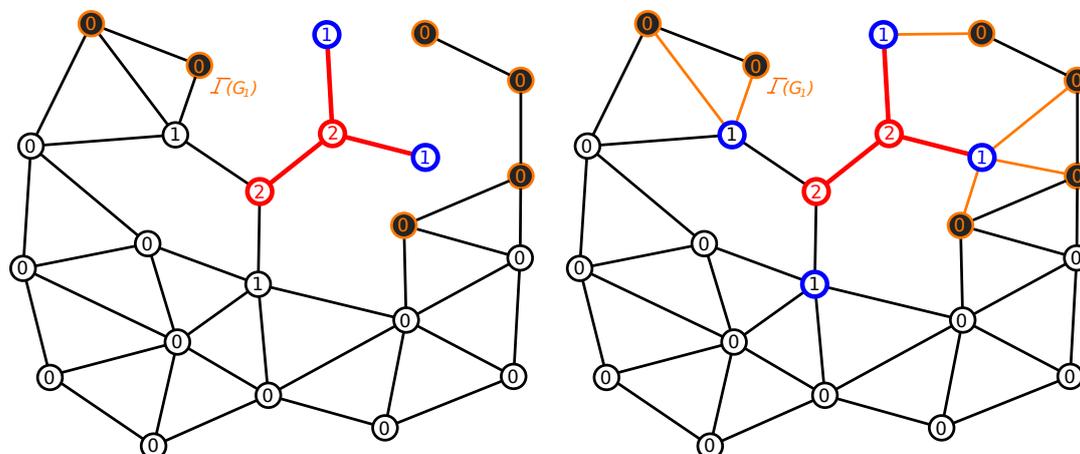


FIGURE 3.5 – Exemple des étapes 1 et 2 de l'intégration d'un motif avec recouvrement.

Comme le montre la Figure 3.6a, le sommet de l'enveloppe de priorité 2 a été fusionné avec le sommet du motif qui le recouvrait. Comme ses arêtes ont été conservées,

le motif est déjà en partie rattaché à l'enveloppe. Dans ce cas la *bordure*( $M$ ) est redéfinie pour s'adapter à la fusion. La nouvelle bordure est représentée sur la Figure 3.6b, ainsi que le rattachement qui en découle.



(a) Suppression de  $G_1$  dans  $S$  après fusion du sommet de priorité 2 avec le sommet le recouvrant dans le motif. (b) Rattachement de  $M$  au reste de l'enveloppe par la nouvelle *bordure*.

FIGURE 3.6 – Exemple des étapes 3 et 4 de l'intégration d'un motif avec recouvrement.

En résumé, les sommets de priorité 1 et 2 en collision avec un motif ont besoin d'être recouverts pour que le motif puisse être intégré. Les sommets de priorité 2 sont alors directement fusionnés au motif et la bordure du motif a besoin d'être redéfinie.

Dans la suite, nous allons détailler les données nécessaires à l'intégration d'un motif. De plus, nous donnerons aussi la procédure permettant de redéfinir la bordure.

## 1.2 Positionnement d'un motif et recouvrement

Afin de pouvoir intégrer un motif liant, plusieurs données sont nécessaires. Pour commencer, il faut savoir où l'intégrer dans l'enveloppe. Ce point de départ est l'un des sommets de l'enveloppe. Il est appelé **sommet d'intégration** du motif. Ce sommet fournit plusieurs données. La première est sa position dans l'espace, ses coordonnées 3D. La seconde est la *direction libre* qui a été utilisée pour le construire. Enfin la dernière est le vecteur *normal* du sommet du substrat dont il est issu (voir chapitre 2 section 3.1).

Les coordonnées du sommet d'intégration servent de coordonnées au sommet, que nous appelons **central**, du motif. Le sommet central du motif est le sommet à partir duquel les coordonnées des autres sommets du motifs sont calculés. Dans le cas des motifs liants, le sommet central du motif est toujours un sommet de priorité 2. La Figure 3.7 montre deux exemples de positionnement d'un motif en utilisant des sommets centraux différents.

La direction libre permet de déterminer la direction opposée au substrat. Elle indique la direction vers laquelle les autres sommets du motif doivent être positionnés

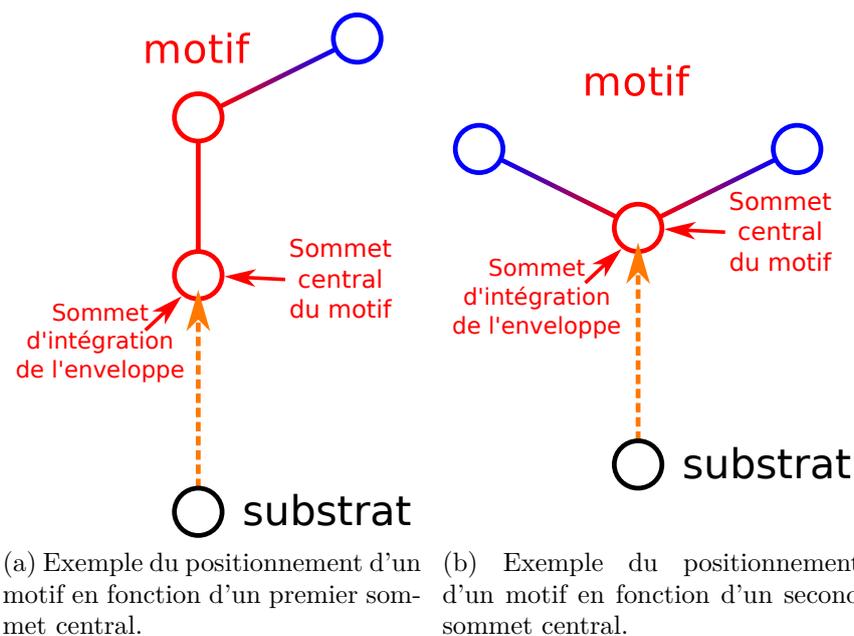


FIGURE 3.7 – Exemples de positionnement d'un motif en fonction de son sommet *central*.

pour ne pas être intégrés dans l'enveloppe. Nous verrons dans la suite qu'elle sert également d'axe de rotation au motif. La Figure 3.8 montre l'exemple de deux motifs positionnés en fonction de la direction libre du sommet d'intégration.

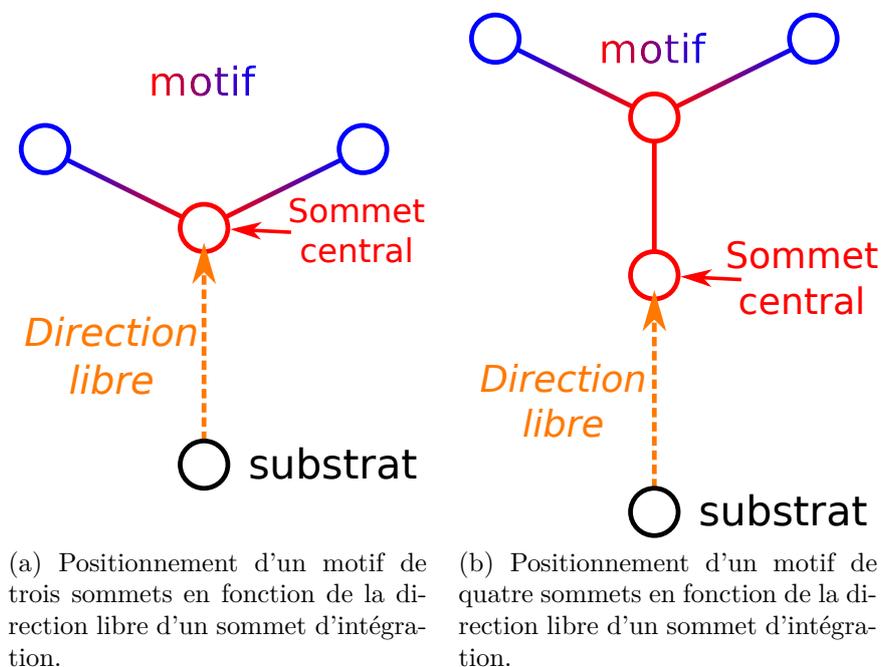


FIGURE 3.8 – Exemple de positionnement de deux motifs différents en fonction de la direction libre d'un sommet d'intégration.

Le vecteur *normal* va ensuite permettre de déterminer l'*orientation* du motif. En

effet, puisque nous sommes dans un espace 3D, un second vecteur est nécessaire pour déterminer l'orientation d'un motif. Si on utilisait seulement la direction libre, le motif pourrait prendre n'importe quelle position autour de l'axe du premier vecteur (qui est ici la direction libre), c'est pourquoi nous utilisons également le vecteur *normal* pour déterminer l'orientation du motif. De plus, en utilisant le vecteur *normal* du sommet du substrat, nous garantissons une cohérence entre le positionnement des sommets du substrat et ceux des solutions générées. La Figure 3.9 montre deux exemples du positionnement d'un motif en fonction d'une direction libre et d'un vecteur *normal*. Dans les deux exemples la direction libre est la même mais le vecteur *normal* est différent d'un exemple à l'autre.

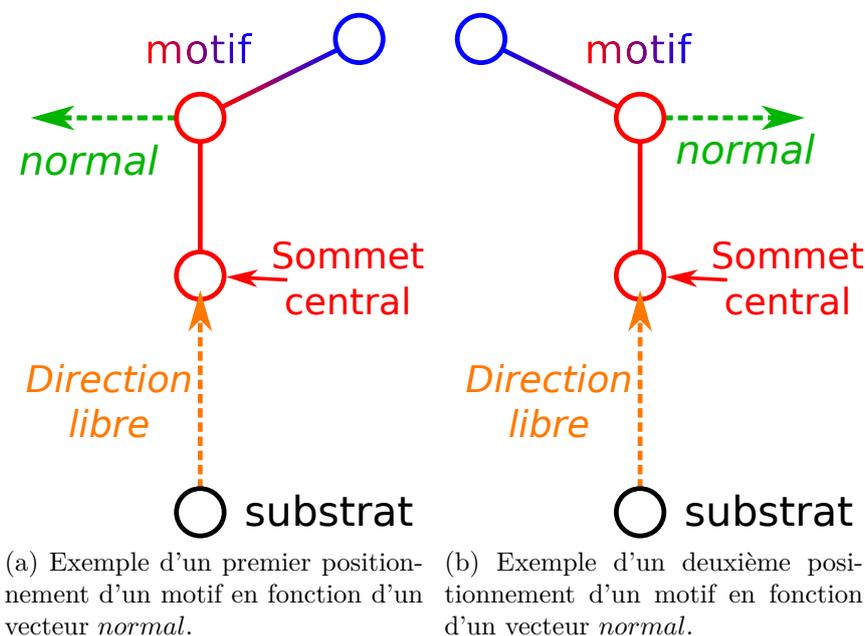


FIGURE 3.9 – Exemples de positionnement d'un même motif en fonction du vecteur *normal* d'un sommet d'intégration.

Cependant, comme nous l'avons expliqué précédemment, il est possible que des sommets de l'enveloppe de priorité 1 et 2 puissent rentrer en collision avec les sommets du motif. Dans ce cas, l'orientation du motif est redéfinie en fonction des sommets en collision. Nous cherchons alors à modifier l'orientation afin que les sommets de priorité 1 et 2 de l'enveloppe soient recouverts par des sommets du motif. Comme nous l'avons vu dans la section 3.1.2 du chapitre 1, il y a une tolérance sur les angles des motifs et la tailles des liaisons. Pour trouver la nouvelle orientation du motif et les positions de ses sommets, nous utilisons les positions des coordonnées des sommets de priorité 1 et 2 en collision, ainsi que les coordonnées du sommet d'intégration comme coordonnées de certains sommets du motif. Si une nouvelle orientation du motif peut être trouvée en recouvrant les sommets de priorité 1 et 2 et en respectant les tolérances, alors le motif est intégré avec les nouvelles positions des sommets du motif. Dans le cas contraire le motif ne sera pas intégré.

### 1.3 Bordure et rattachement

Le rattachement du motif au reste de l'enveloppe se fait en deux étapes. La première consiste à chercher les sommets situés sur la bordure du motif et la seconde est de rattacher les sommets de l'ensemble  $\Gamma(G_1)$  aux sommets de cette bordure.

Si le motif ne subit aucune modification au cours de l'intégration la bordure est l'ensemble des sommets de priorité 1 du motif. Cependant, comme nous l'avons dit précédemment, le motif peut être modifié au cours de l'intégration si un sommet de priorité 2 est recouvert par l'un des sommets du motif. Si un sommet de priorité 2 fusionne avec un sommet du motif, cela signifie que deux motifs (ou plus) ont été rattachés ensemble. Un ajustement de la bordure est alors nécessaire, comme dans l'exemple de la Figure 3.6b.

Afin de trouver la nouvelle bordure, on utilise l'algorithme 6. Dans cet algorithme, *traité* est l'ensemble des sommets déjà visités. Au départ, les ensembles *traité* et *bordure* sont vides et l'identifiant du sommet de départ est celui du sommet central du motif intégré. À la fin de l'algorithme les sommets de *bordure* sont tous les sommets de priorité 1 qui entourent les sommets de priorité 2 de l'agglomérat de motif.

---

#### Algorithme 6 : RechercheBordure

---

**Entrées :** Graphe de l'enveloppe  $S$ , Ensemble *traité*, Identifiant *id* d'un sommet

**Sorties :** Ensemble *bordure* contenant les sommets de la bordure

```

1 traité = traité ∪ id
2 si prioritéid == 1 alors
3   |   retourner bordure = bordure ∪ id
4 fin
5 pour tous les voisins v de id faire
6   |   si v ∉ traité alors
7     |   |   bordure = bordure ∪ RechercheBordure( $S$ , traité, v)
8     |   fin
9 fin
10 retourner bordure

```

---

Après avoir défini l'ensemble des sommets de la bordure, pour chaque sommet de l'ensemble  $\Gamma(G_1)$ , nous cherchons le sommet de la bordure le plus proche de lui géométriquement. Puis nous ajoutons une arête dans l'enveloppe  $S$  entre ce sommet de  $\Gamma(G_1)$  et le sommet de la *bordure* le plus proche.

## 2 Construction des liaisons aromatiques

Nous venons d'expliquer la méthode d'intégration des motifs de manière générale. Ici, nous nous intéressons plus particulièrement à l'intégration des motifs aromatiques. Dans cette partie, nous allons décrire la méthode nous permettant de déterminer quels sont les sommets de l'enveloppe qui peuvent servir de sommets d'intégration pour les motifs aromatiques. De plus, nous montrerons que, bien qu'il existe plusieurs motifs aromatiques, nous pouvons les reconstruire en intégrant successivement un seul et unique motif qui est le motif triangulaire utilisé dans les exemples de la partie précé-

dente. Nous verrons également que l'ordre dans lequel les motifs sont insérés n'a pas d'importance.

Les motifs aromatiques sont les premiers à être intégrés dans l'enveloppe. Pour rappel, les composés aromatiques sont des ensembles d'atomes formant des cycles plans. Ce sont donc des cycles composés d'atomes à géométrie triangulaire. Lorsque deux cycles aromatiques interagissent entre eux, ils forment ce qu'on appelle ici des « liaisons aromatiques ». En intégrant des motifs aromatiques dans l'enveloppe, on cherche à garantir la formation de liaisons aromatiques entre le substrat et les molécules cages.

Dans le but d'intégrer des motifs aromatiques, la première étape est de trouver les sommets d'intégration de ces motifs. Comme on a pu le voir dans le chapitre précédent, lorsqu'un cycle composé de sommets à géométrie triangulaire est présent dans le substrat, deux cycles similaires apparaissent dans l'enveloppe. Ces cycles sont la conséquence des translations effectuées dans les directions libres perpendiculaires au plan du cycle comme on le rappelle sur la Figure 3.10. Les sommets d'intégration des motifs aromatiques sont les sommets de ses cycles de l'enveloppe.

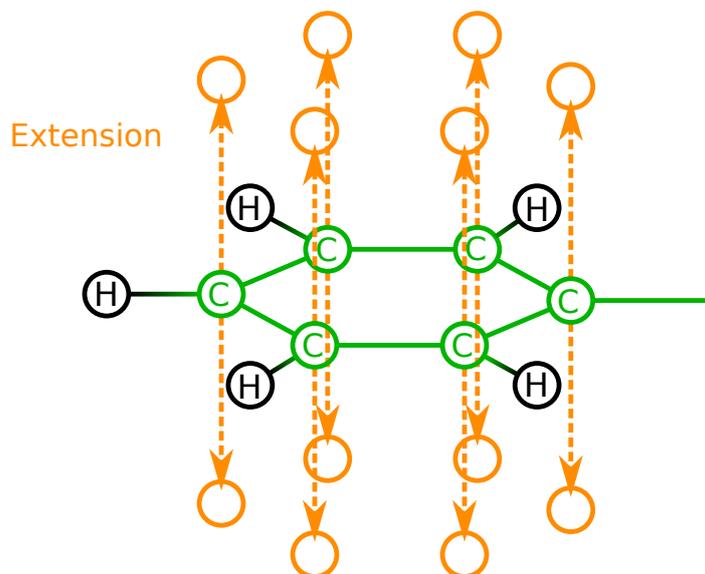


FIGURE 3.10 – Exemple d'extension d'un composé aromatique

Il arrive cependant que plusieurs de ces sommets ne soient pas conservés lors de la formation finale de l'enveloppe. Dans ce cas les cycles aromatiques de l'enveloppe sont considérés comme *partiels*, c'est-à-dire qu'il manque une partie du cycle. Les cycles aromatiques partiels sont donc les arcs de cercle de l'enveloppe dont les sommets sont issus des directions libres perpendiculaires des sommets du substrat appartenant à des cycles aromatiques. Les liaisons entre les cycles partiels seront moins fortes que celles des cycles complets mais cela n'empêchera pas une bonne interaction entre le substrat et les cages moléculaires à cet endroit-là.

La Figure 3.11, nous montre les deux motifs aromatiques complets pouvant être présents dans l'enveloppe. Chacun des deux pouvant donner lieu à plusieurs motifs partiels en fonction du nombre de sommets restants dans l'enveloppe après la fin de sa construction. Après intégration des motifs, les sommets des cycles aromatiques sont de priorité 2 et les autres sommets des motifs sont de priorité 1 afin de pouvoir rattacher

les motifs au reste de l'enveloppe. En d'autres termes, les sommets d'intégration des cycles aromatiques sont recouverts par les sommets de priorité 2 des motifs aromatiques.

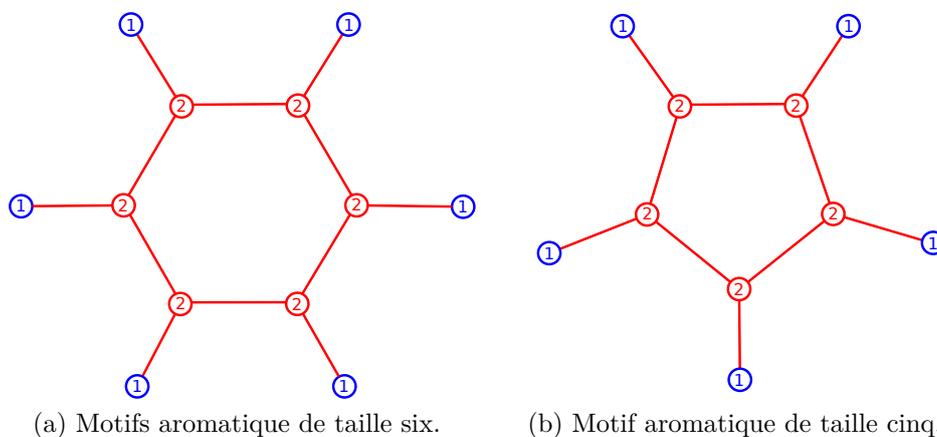


FIGURE 3.11 – Motifs aromatiques complets.

À partir de l'ensemble des sommets d'intégration des motifs aromatiques, il est difficile de déterminer les sommets appartenant à un même motif aromatique dans l'enveloppe. Ainsi, plutôt que d'essayer de recouvrir plusieurs sommets d'intégration de l'enveloppe par un motif aromatique, on va utiliser le fait que tous les sommets des cycles des motifs aromatiques soient des sommets à géométrie triangulaire.

Pour chaque sommet d'intégration on va intégrer un motif triangulaire dans l'enveloppe. Par exemple, un motif aromatique complet de taille six, sera la conséquence de l'intégration de six motifs triangulaires. Le problème ici est que la direction libre des sommets d'intégration est la même que le vecteur *normal*. C'est pourquoi, la méthode de recouvrement est utilisée pour déterminer la position des sommets des motifs triangulaires. Puisque ces cycles de l'enveloppe sont des projections des cycles du substrat, les distances entre les sommets des cycles sont conservées de même que les angles. Dans le cas des cycles complets, pour chaque sommet d'intégration, on doit pouvoir trouver deux sommets voisins pouvant être recouverts lors de l'intégration d'un motif triangulaire. Dans le cas des cycles partiels, au moins un sommet voisin doit pouvoir être recouvert par l'un des sommets du motif (pour les extrémités des arcs).

La Figure 3.12 montre l'exemple des intégrations successives de motifs triangulaires pour recouvrir un cycle aromatique de taille six. Les sommets du cycle sont utilisés les uns après les autres comme sommet d'intégration. Quand ils sont utilisés comme sommet d'intégration, ils sont recouverts par le sommet de priorité 2 qui est le sommet central du motif. Lorsque leurs voisins sont utilisés comme sommet d'intégration, ils sont recouverts par l'un des sommets de priorité 1 des motifs. Comme nous l'avons vu précédemment, s'ils étaient déjà de priorité 2, ils conservent cette priorité. Au final, peu importe l'ordre dans lequel ils sont utilisés comme sommet d'intégration, le résultat sera le même puisqu'ils auront tous été recouverts par un sommet de priorité 2. En utilisant les intégrations successives de motifs triangulaires, nous obtenons bien l'intégration d'un motif aromatique recouvrant tous les sommets du cycle.

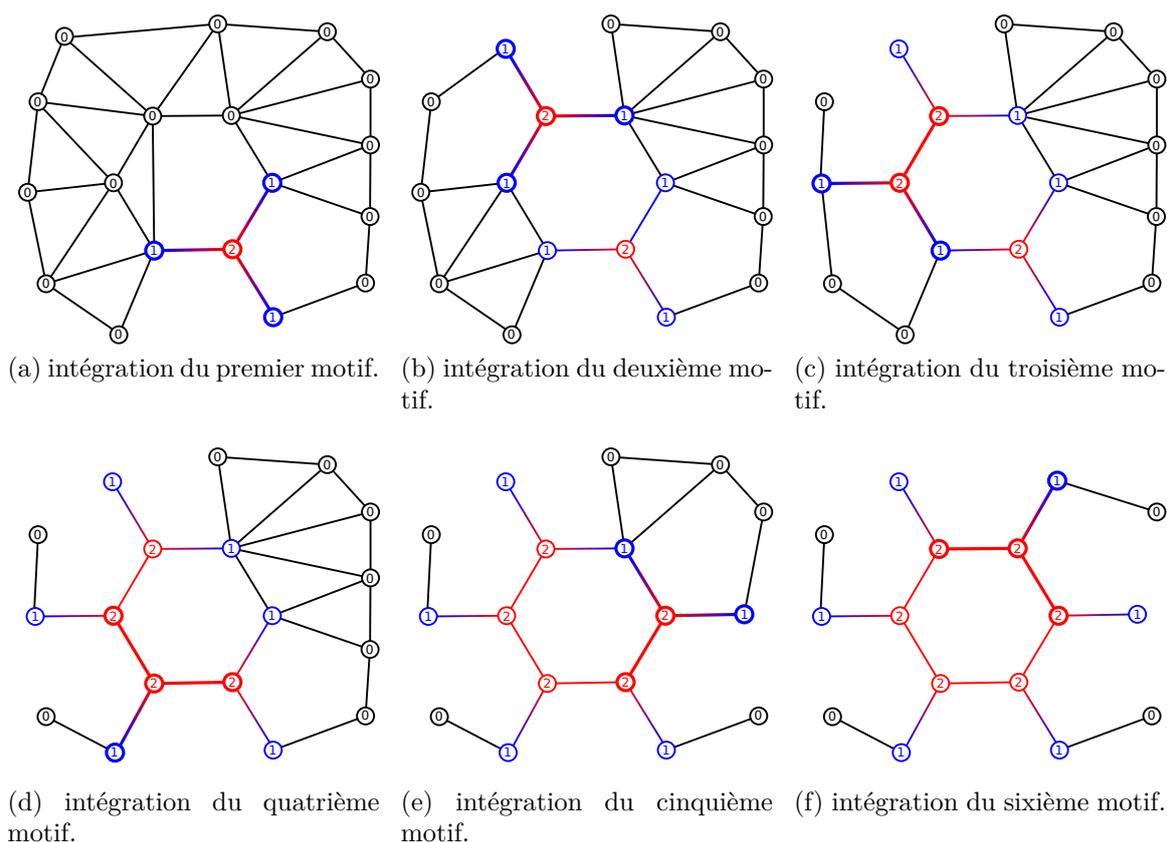


FIGURE 3.12 – intégration d'un cycle complet.

En résumé, les motifs aromatiques sont les premiers insérés dans l'enveloppe. Les sommets d'intégration de ses motifs sont ceux issus des directions libres perpendiculaires des sommets des cycles aromatiques du substrat. De plus, plutôt que d'intégrer un motif aromatique en remplaçant plusieurs sommets d'intégration en même temps, nous intégrons un motif triangulaire sur chaque sommet d'intégration de l'enveloppe de manière successive. Les motifs aromatiques sont ainsi intégrés grâce à la fusion de plusieurs motifs triangulaires.

### 3 Construction des Liaisons Hydrogènes

Nous allons maintenant étudier plus particulièrement l'intégration des motifs hydrogènes et les sommets pouvant servir de sommets d'intégration pour ces motifs. Nous verrons qu'il existe plusieurs types de motifs hydrogènes, et qu'un choix doit être effectué parmi eux pour chaque sommet d'intégration utilisé. De plus, nous montrerons que tous les sommets d'intégration ne peuvent pas forcément être utilisés dans une même solution. Chaque ensemble de sommets d'intégration apportera un guide différent pour les cages moléculaires.

Comme pour les motifs aromatiques, pour garder une cohérence chimique, il n'est pas possible d'intégrer des motifs hydrogènes de façon aléatoire dans l'enveloppe. Pour qu'un sommet puisse être un sommet d'intégration pour un motif hydrogène, il doit être issu d'un sommet du substrat pouvant être impliqué dans un liaison hydrogène. De plus, les sommets d'un motif hydrogène ne peuvent pas être impliqués dans plu-

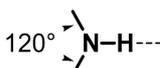
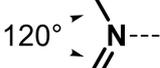
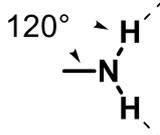
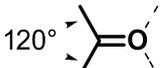
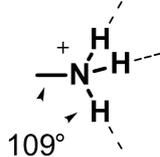
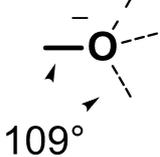
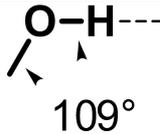
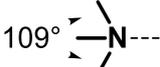
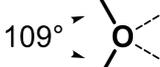
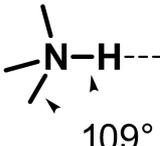
sieurs liaisons hydrogènes en même temps. Ainsi plusieurs cages moléculaires différentes peuvent être construites en fonction des sommets de l'enveloppe qui ont été remplacés par des motifs hydrogènes.

Dans un premier temps, nous allons donner un aperçu des différents motifs hydrogènes. Puis nous expliquerons comment déterminer les sommets pouvant servir de sommet d'intégration. Pour finir nous décrirons la méthode utilisée pour déterminer le choix du motif à insérer pour chaque sommet d'intégration utilisé, ainsi que la manière dont les différentes solutions sont générées à partir de l'enveloppe.

### 3.1 Donneur / Accepteur

Pour rappel, pour qu'une liaison hydrogène se forme, on a besoin d'un donneur et d'un accepteur. Un donneur est un ensemble d'atomes composé d'au moins un atome hydrogène relié à un hétéro-atome. Un hétéro-atome est un atome qui n'est ni un carbone ni un hydrogène. Un accepteur est, quant à lui, composé au minimum d'un atome d'azote ou d'oxygène (ou de fluor) possédant au moins un doublet non-liant. Le tableau 3.1 récapitule les différents motifs moléculaires donneurs et accepteurs qu'il est possible de trouver en chimie organique.

TABLE 3.1 – Modifs moléculaires des liaisons hydrogènes

Monoliasion		Diliasion		Triliasion	
Donneur	Accepteur	Donneur	Accepteur	Donneur	Accepteur
					
					
					

En dehors du côté donneur/accepteur, il existe trois types de motifs. Les monoliasions, les diliasions et les triliasions. Comme énoncé plus haut, lorsqu'un atome d'un motif hydrogène est lié par une liaison hydrogène à un atome d'un motif complémentaire, il est difficile pour les autres atomes du motif d'être également impliqués dans des liaisons hydrogènes en même temps. Ainsi, même si, en théorie, plusieurs sommets d'un motif diliasion ou triliasion peuvent être impliqués dans des liaisons hydrogènes, en pratique si l'un des sommets est impliqué dans une liaison hydrogène, les autres ne le seront pas. C'est pourquoi nous construisons ce que nous appelons des graphes de dépendance afin de déterminer quels sont les sommets ne pouvant pas être impliqués en même temps dans des liaisons hydrogènes.

## 3.2 Graphes des dépendances

Il existe deux types de graphes de dépendance. Le premier est le **graphe de dépendance du substrat**. Ce graphe est composé des sommets du substrat pouvant être impliqués dans des liaisons hydrogènes, c'est-à-dire les sommets d'hydrogène reliés à des hétéro-atomes et les sommets d'oxygène, d'azote ou de fluor possédant au moins un doublet non-liant. Il existe une arête entre deux sommets dans ce graphe, si les deux sommets ne peuvent pas être impliqués en même temps dans une liaison hydrogène. Une arête est donc présente entre deux sommets d'un même motif hydrogène.

Le deuxième est le **graphe de dépendance de l'enveloppe**. Ce graphe est composé des sommets pouvant servir de sommet d'intégration pour les motifs hydrogènes. Ce sont les sommets de l'enveloppe situés dans l'une des directions libres de l'un des sommets du graphe de dépendance du substrat. Cependant dans le cas des sommets à géométrie triangulaire, seuls les sommets issus des directions libres des doublets non-liants peuvent servir de sommets d'intégration pour des motifs hydrogènes. De même que pour le graphe de dépendance, il existe une arête entre deux sommets dans ce graphe, si les deux sommets ne peuvent pas être impliqués en même temps dans une liaison hydrogène.

Dans la suite, nous allons détailler la manière de déterminer, pour chacun de ces graphes de dépendance, s'il existe une dépendance (une arête) entre deux sommets.

### 3.2.1 Graphe de dépendance du substrat

On sait que les donneurs sont les atomes d'hydrogènes reliés à un atome autre qu'un carbone et que les accepteurs sont les atomes d'azote, d'oxygène possédant au moins un doublet non-liant.

Il existe deux cas dans lesquels deux sommets du substrat peuvent être dépendants. Le premier est le cas où plusieurs sommets d'hydrogène sont reliés à un même hétéro-atome, c'est-à-dire les sommets d'hydrogène des motifs donneurs diliaisons et triliaisons. Puisque l'hétéro-atome ne peut être impliqué que dans une liaison à la fois, si l'un de ses sommets hydrogènes a déjà établi une liaison, les autres ne pourront plus en établir. Il y a donc des arêtes entre les sommets des différents hydrogènes reliés à un même hétéro-atome dans le graphe de dépendance du substrat.

Le deuxième cas est celui où un sommet hydrogène est relié à un sommet d'oxygène ou d'azote. Ici, le sommet d'hydrogène peut être utilisé comme un donneur et le sommet d'oxygène ou d'azote peut être utilisé comme un accepteur. Cependant si l'un est utilisé comme donneur l'autre ne pourra plus être utilisé comme accepteur, et inversement. Il y a donc une dépendance entre le sommet d'hydrogène et le sommet auquel il est relié dans le substrat.

L'algorithme 7 construit le graphe de dépendance du substrat. Pour chaque sommet du substrat qui ne représente ni un atome de carbone, ni un atome d'hydrogène, on regarde s'il peut faire partie d'un motif donneur ou accepteur. Si ce sommet représente un atome d'oxygène ou d'azote et qu'il possède au moins un doublet non-liant, alors on le rajoute au graphe de dépendance puisqu'il s'agit d'un accepteur. Ensuite on rajoute au graphe de dépendance tous ses voisins représentant des atomes hydrogènes qui sont des donneurs. Et pour finir on ajoute une arête entre chaque couple de

sommets venant d'être ajoutés. Dans la suite, nous appelons le graphe de dépendance du substrat  $D_G$ , afin de ne pas le confondre avec celui de l'enveloppe.

---

**Algorithme 7 : DépendanceSubstrat**


---

**Entrées :** Substrat  $G$

**Sorties :** Graphe  $D_G$

```

1 pour tous les sommets  $v$  de  $V_G$  faire
2   si  $\text{symbole}(v) \neq H$  et  $\text{symbole}(v) \neq C$  alors
3      $l = \emptyset$  si  $(m_v > 0)$  et  $(\text{symbole}(v) = O \text{ ou } N \text{ ou } F)$  alors
4        $l = l \cup v$ 
5        $V_{D_G} = V_{D_G} \cup v$ 
6     fin
7     pour tous les voisins  $n_i$  de  $v$  faire
8       si  $\text{symbol}(n_i) = H$  alors
9          $l = l \cup n_i$ 
10         $V_{D_G} = V_{D_G} \cup n_i$ 
11      fin
12    fin
13    pour tous les couples de sommets  $(u, v)$  de  $l$  faire
14      Ajouter l'arête  $(u, v)$  à  $E_{D_G}$ .
15    fin
16  fin
17 fin
18 retourner  $D_G$ 

```

---

### 3.2.2 Graphe de dépendance de l'enveloppe

À partir du graphe de dépendance du substrat  $D_G$ , on peut construire le graphe de dépendance de l'enveloppe  $D_S$ . Lors de l'expansion, on ajoute au graphe de dépendance de l'enveloppe tous les sommets issus des directions libres des doublets non-liants des sommets appartenant au graphe de dépendance du substrat. C'est pourquoi même si des sommets à géométrie triangulaire font partie du graphe de dépendance du substrat, les sommets de l'enveloppe issus des directions libres perpendiculaires au plan du triangle ne sont pas rajoutés au graphe de dépendance de l'enveloppe.

Une arête est ensuite ajoutée entre deux sommets du graphe  $D_S$  s'ils sont issus du même sommet du substrat ou s'il existe une arête dans le graphe de dépendance du substrat entre les deux sommets dont ils sont issus.

La Figure 3.13 montre les graphes de dépendance associés à un substrat et son enveloppe. Le substrat représenté sur la Figure 3.13a possède trois sommets pouvant être impliqués dans des liaisons hydrogènes, le numéro 5, qui représente un atome d'azote, le sommet 9, qui représente un atome d'oxygène et le sommet 10, qui représente un atome d'hydrogène relié à un hétéro-atome. Le graphe de dépendance de ce substrat sera donc composé de trois sommets. Et étant donné que le sommet 9 et le sommet 10 sont liés dans le substrat, ils ne pourront pas être impliqués en même temps dans des liaisons hydrogènes. Une arête est donc intégrée dans le graphe de dépendance entre ces deux sommets. Pour ce substrat, on obtient donc le graphe de dépendance de la Figure 3.13b.

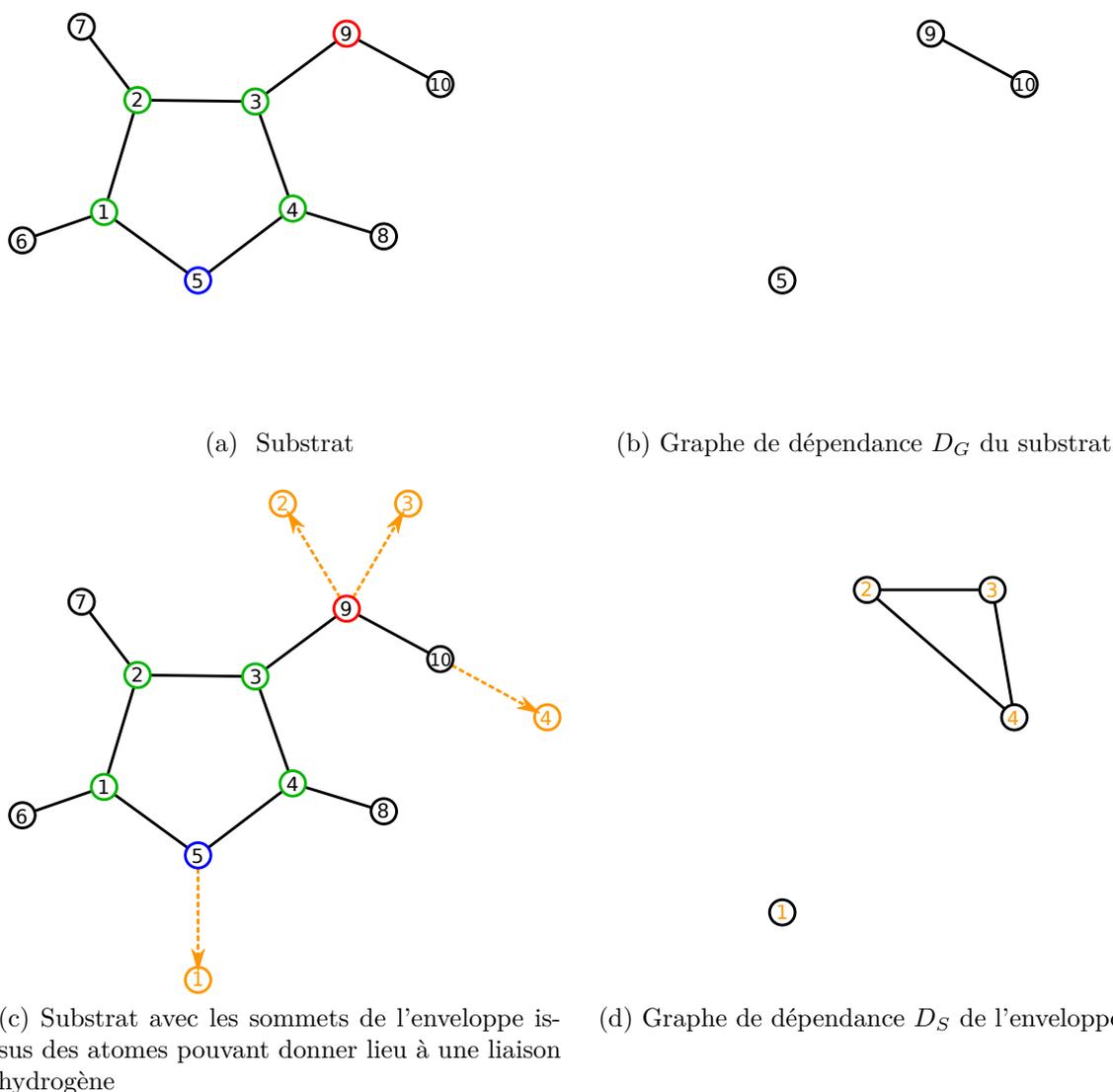


FIGURE 3.13 – Graphes de dépendance.

Le sommet 5 représente un atome d'azote de topologie  $(2, 1)$ , il donnera donc un sommet dans l'enveloppe pouvant servir de sommet d'intégration. Le sommet 9 représente un atome d'oxygène de topologie  $(2, 2)$ , il possède donc deux doublets non-liants et donnera par conséquent deux sommets d'intégration dans l'enveloppe. Et enfin le sommet d'hydrogène, c'est-à-dire le sommet 10, donnera lieu à un sommet dans l'enveloppe. La Figure 3.13c est une représentation du substrat avec ces quatre sommets de l'enveloppe issus des sommets du substrat pouvant être impliqués dans des liaisons hydrogènes.

Puisqu'ils sont issus de sommets du substrat pouvant être impliqués dans des liaisons hydrogènes, ces quatre sommets sont inclus dans le graphe de dépendance  $D_S$  de l'enveloppe représenté sur la Figure 3.13d. Dans la mesure où les sommets 2 et 3 sont issus du même sommet du substrat, une arête est ajoutée entre les deux dans le graphe

de dépendance de l'enveloppe. De plus, puisqu'il existe une arête entre les sommets 9 et 10 dans le graphe de dépendance du substrat, il y a également des arêtes entre les sommets de l'enveloppe issus du sommet 9 et le sommet de l'enveloppe issu du sommet 10.

Dans cet exemple, il y a donc quatre sommets de l'enveloppe pouvant servir de sommet d'intégration pour des motifs hydrogènes. Cependant puisque plusieurs d'entre eux sont reliés, ils ne pourront pas tous être utilisés dans les mêmes solutions.

### 3.3 Intégration des liaisons hydrogènes

La dernière étape est d'intégrer les motifs hydrogènes en utilisant les sommets du graphe de dépendance  $D_S$  comme sommet d'intégration.

#### 3.3.1 Choix du motif à intégrer

Comme décrit précédemment, il existe plusieurs motifs hydrogène. Il est nécessaire de faire un choix parmi ces motifs. Pour commencer, il faut établir si on insère un motif donneur ou un motif accepteur. Pour qu'une liaison hydrogène s'établisse les deux motifs sont nécessaires. Par conséquent, si le motif du substrat est un donneur, on intégrera dans l'enveloppe un motif accepteur et inversement, si le motif du substrat est un accepteur, on intégrera un donneur. Si le sommet d'intégration de l'enveloppe est issu d'un sommet du substrat représentant un hydrogène (donneur), nous intégrons un motif accepteur et si le sommet de l'enveloppe est issu d'un sommet du substrat représentant un oxygène ou un azote (accepteur), nous intégrons un motif donneur.

Cependant il existe plusieurs motifs donneurs et plusieurs motifs accepteurs. Nous avons fait un choix parmi ces motifs. Dans la mesure où les motifs monoliasions sont les plus courants, nous avons pris le parti de n'intégrer que ces motifs. Comme nous l'avons vu dans le tableau 3.1, il y a un motif donneur et un motif accepteur à géométrie triangulaire, et un motif donneur et un motif accepteur à géométrie tétraédrique. Puisque nous utilisons comme vecteur *normal* du sommet d'intégration le vecteur *normal* du sommet du substrat dont il est issu, nous utilisons également la géométrie de ce sommet du substrat pour déterminer la géométrie du motif à intégrer.

Dans le cas où le sommet du substrat est un accepteur à géométrie triangulaire, nous intégrons le motif donneur à géométrie triangulaire et dans le cas où c'est un accepteur à géométrie tétraédrique nous utilisons le donneur à géométrie tétraédrique. De même, si le sommet du substrat est un donneur, par conséquent un sommet d'hydrogène, nous utilisons le vecteur *normal* et la géométrie du sommet auquel il est relié dans le substrat pour déterminer si nous intégrons le motif accepteur à géométrie triangulaire ou celui à géométrie tétraédrique.

#### 3.3.2 Choix des sommets à remplacer parmi ceux du graphe de dépendance

Comme expliqué précédemment, les sommets qui sont reliés dans le graphe de dépendance de l'enveloppe ne peuvent pas servir comme sommet d'intégration dans une même solution. C'est pourquoi à partir d'une même enveloppe, on va pouvoir générer plusieurs solutions.

Plusieurs copies de l'enveloppe vont être générées de telle sorte que chaque copie ne conserve pas les mêmes sommets dans leur graphe de dépendance. Le but est de retirer les dépendances en ne conservant pour chaque copie qu'un seul des sommets des cycles du graphe de dépendance. Pour chaque arête  $[u, v]$  d'un graphe de dépendance  $S_i$ , une copie de ce graphe est créée de telle sorte que lui conserve le sommet  $u$  et sa copie conserve le sommet  $v$ .

---

**Algorithme 8 : Génération des Graphes sans dépendance**


---

**Entrées :** Graphes de dépendance  $D_{S_0}$  du graphe  $S_0$

**Sorties :** Graphes  $D_{S_i}$

```

1 pour tous les graphes  $S_i$  faire
2   | pour tous les arêtes  $[u, v]$  de  $D_{S_i}$  faire
3   |   |  $S_j =$  copie  $S_i$ 
4   |   | Supprimer  $v$  de  $S_i$ 
5   |   | Supprimer  $u$  de  $S_j$ 
6   |   fin
7 fin

```

---

En utilisant cet algorithme plusieurs copies peuvent avoir conservé les mêmes sommets d'intégration dans leur graphe de dépendance. On supprime alors les solutions identiques.

La Figure 3.2 est un exemple de l'exécution de l'algorithme 8 sur le graphe de dépendance de l'enveloppe trouvé dans la Figure 3.13. Dans cet exemple, le graphe de dépendance de  $S_0$ , c'est-à-dire le graphe de dépendance  $D_{S_0}$ , est composé, au départ, de quatre sommets et trois arêtes.

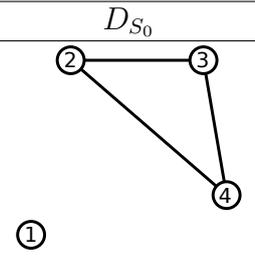
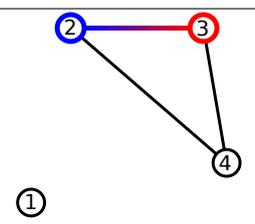
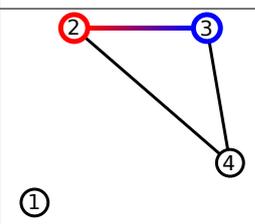
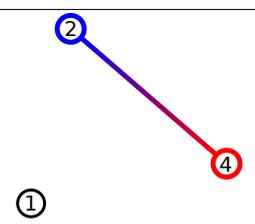
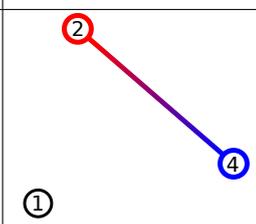
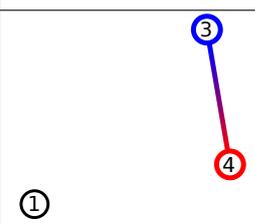
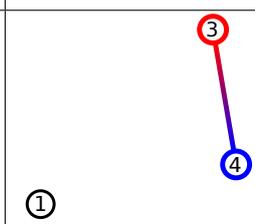
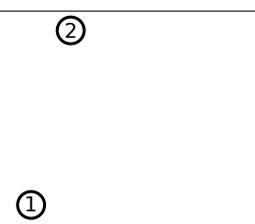
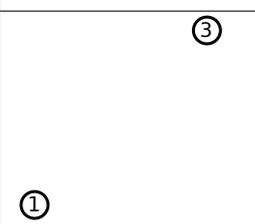
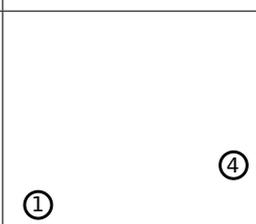
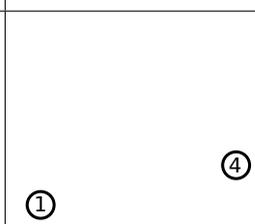
Les arêtes sont retirées les unes après les autres en créant des copies du graphe. À la première étape, l'arête  $[2, 3]$  du graphe  $D_{S_0}$  est traitée. Un nouveau graphe  $S_1$  copie de  $S_0$  est créé. Le sommet 3 est retiré du graphe  $D_{S_0}$  et le sommet 2 est retiré du graphe  $D_{S_1}$ . Puis l'arête  $[2, 4]$  du graphe  $D_{S_0}$  est traité, ce qui a pour conséquence la création un nouveau graphe  $S_2$ , copie du graphe  $S_0$ , dont le graphe de dépendance  $D_{S_2}$  est privé du sommet 2, et le retrait du sommet 4 de  $D_{S_0}$ . Comme toutes les arêtes de  $D_{S_0}$  ont été traitées, on passe au graphe suivant qui est  $S_1$ . Dans le graphe de dépendance de  $S_1$ , il ne reste que l'arête  $[3, 4]$ . Un nouveau graphe  $S_3$  est créé, copie du graphe  $S_1$ . Le sommet 3 de  $D_{S_3}$  et le sommet 4 de  $D_{S_1}$ .

À la fin de l'algorithme, les graphes de dépendance n'ont plus d'arête, il n'y a donc plus de dépendance entre les sommets d'intégration qui ont été conservés. Cependant plusieurs copies ( $D_{S_1}$  et  $D_{S_4}$ ) ont le même graphe de dépendance final. Dans ce cas, les doublons sont supprimés. Ainsi dans notre exemple, seules les trois copies sont conservées. Chaque copie donnera lieu à une solution différente après intégration des motifs hydrogènes.

## 4 Conclusion

Dans cette étape, nous commençons par intégrer les motifs aromatiques puis nous intégrons les motifs hydrogènes. Les motifs hydrogènes sont intégrés après car, contrairement aux motifs aromatiques, leur intégration peut générer plusieurs solutions.

TABLE 3.2 – Exemple d'application de l'algorithme 8

	$D_{S_0}$	$D_{S_1}$	$D_{S_2}$	$D_{S_3}$
Départ				
Étape 1				
Étape 2				
Étape 3				
Arrivée				

Les motifs aromatiques sont intégrés sur les sommets de l'enveloppe provenant des sommets des cycles aromatiques du substrat. Au lieu d'intégrer directement des motifs aromatiques, nous utilisons le fait que tous les motifs aromatiques sont l'agglomération de plusieurs motifs triangulaires pour n'intégrer que des motifs triangulaires. L'intégration successive de motifs triangulaires sur tous les sommets d'intégration possibles permet de reconstruire les motifs aromatiques.

Les motifs hydrogènes sont intégrés sur les sommets de l'enveloppe provenant des sommets du substrat appartenant eux-mêmes à un motif hydrogène. Cependant, certains sommets d'intégration ne peuvent pas être utilisés dans une même solution. Pour chaque ensemble de sommets d'intégration pouvant être utilisés ensemble pour intégrer des motifs hydrogènes, nous créons une solution. Le choix du motif hydrogène intégré sur un sommet de l'enveloppe, dépend du sommet du substrat dont il est issu. Si le sommet du substrat est un sommet d'hydrogène, alors le motif intégré est un motif de type donneur, sinon c'est un motif de type accepteur. Parmi les motifs accepteurs, nous intégrons le motif qui a la même géométrie que le sommet du substrat. Il en est

de même pour les motifs donneurs intégrés. Le motif intégré est celui ayant la même géométrie que le sommet du substrat auquel est rattaché le sommet d'hydrogène dont est issu le sommet d'intégration.

Enfin l'intégration d'un motif s'effectue en quatre étapes. Nous commençons par déterminer quels sont les sommets de l'enveloppe qui doivent être retirés pour permettre l'intégration du motif. Ensuite, nous déterminons l'ensemble des voisins des sommets de cette ensemble. Puis nous supprimons les sommets à retirer. Et enfin, nous connectons les sommets de la bordure du motif aux anciens voisins des sommets supprimés. Dans le cas où les sommets à remplacer sont des sommets de priorité 2 ou 1, les sommets doivent être recouverts par des sommets du motif pour que celui-ci puisse être inséré.

Les solutions ainsi générées permettent d'établir des guides pour la construction de cages moléculaires pouvant capturer le substrat donné.

# Chapitre 4

## Etude de cas

---

1	Description de l'application . . . . .	64
2	Étude de l'approche sur la molécule l'adénosine . . . . .	64
3	Étude d'approche sur la molécule de saccharose . . . . .	73
4	Étude de l'approche sur les molécules d'acétanilide et D-tyrosine . . . . .	76
5	Conclusion . . . . .	83

---

*Dans ce chapitre, nous montrons des résultats de l'approche que nous avons présenté jusqu'à maintenant sur des cas réels. Les exemples que nous utilisons sont tous extraits de la Cambridge Structural Database ([Centre](#)) disponible en ligne. Cette base de données regroupe actuellement les diagrammes de diffractions aux rayons X de plus de 900 000 molécules et est en constante évolution. Afin qu'une molécule de la base de données puisse être utilisée comme substrat, deux conditions sont à remplir. La première est que la molécule possède plus de deux atomes. En effet, si la molécule possède seulement un ou deux atomes, l'approche ne permet pas de pouvoir trouver la topologie de ses atomes (à l'exception de quelques cas bien précis). La deuxième condition est que le graphe moléculaire qui représente la molécule soit connexe.*

*Pour rappel, l'approche que nous présentons part d'une molécule cible que nous appelons substrat. Le but est de générer à partir de ce substrat des guides permettant la construction de molécules cages qui seront capables de le reconnaître et de le capturer. La première phase de l'approche construit la forme idéale autour de laquelle les cages doivent être générées afin d'être les plus spécifiques possible au substrat. C'est ce que nous appelons l'enveloppe. Bien que les molécules cages ne pourront pas avoir exactement cette forme, plus la forme finale des cages est proche de celle-ci, plus la reconnaissance du substrat est précise. La seconde phase permet de positionner autour de la cible des groupements molécules qui permettront aux cages de pouvoir capturer la cible en créant des interactions avec elle.*

*Dans la première partie de ce chapitre, nous décrivons l'application réalisée afin de tester l'approche que nous proposons. Dans la partie suivante, nous détaillons un premier cas d'étude qui est la molécule d'adénosine. Elle a la particularité de mettre en avant tous les aspects de l'approche. Dans la troisième partie nous étudions la molécule de saccharose qui a la caractéristique de pouvoir créer un grand nombre de liaisons hydrogènes. Enfin nous finissons par l'étude de l'impact d'approche en fonction de la géométrie des substrats en utilisant l'exemple des molécules d'acétanilide et de D-tyrosine.*

## 1 Description de l'application

Afin de pouvoir tester l'approche que nous proposons, nous avons réalisé une application. Cette application permet la construction de l'enveloppe d'un substrat, ainsi que la génération de différents guides de construction pour des molécules cages. Ces guides comprennent les motifs liants intéressants pour la construction des cages ainsi que leur position dans l'espace autour du substrat. Cette application a été réalisée en langage *C* et utilise également la bibliothèque *alphashape3d* du langage *R* ([CRAN](#)). Cette application comprend environ 3000 lignes de code réparties sur 5 modules.

Le premier module est le module d'initiation. Il permet de récupérer les données du substrat à partir d'un fichier. Dans le cas où le fichier ne contient que les données des atomes du substrat, il trouve les liaisons en utilisant les rayons de covalence théorique des atomes qui le composent. Ce module calcule également la topologie des sommets du substrat et son graphe de dépendance après extraction des données.

Le deuxième module permet la construction de l'enveloppe. Il calcule les directions libres de chaque sommet afin de déterminer les sommets de l'enveloppe et leur position dans l'espace. De plus, il construit également le graphe de dépendance de l'enveloppe au cours de l'expansion. Puis il utilise la bibliothèque *alphashape3d* implémenté en *R* pour trouver l'enveloppe concave à partir des sommets de l'enveloppe et ainsi construire les arêtes de l'enveloppe.

Le troisième module insère les motifs liants. Il commence par insérer les motifs aromatiques. Puis, il utilise le graphe de dépendance de l'enveloppe pour générer plusieurs solutions et insérer les motifs hydrogènes.

Le quatrième module gère les structures de données utilisées pour stocker les différents graphes. Les graphes sont représentés par des tableaux de sommets, et chaque sommet possède la liste de ses voisins.

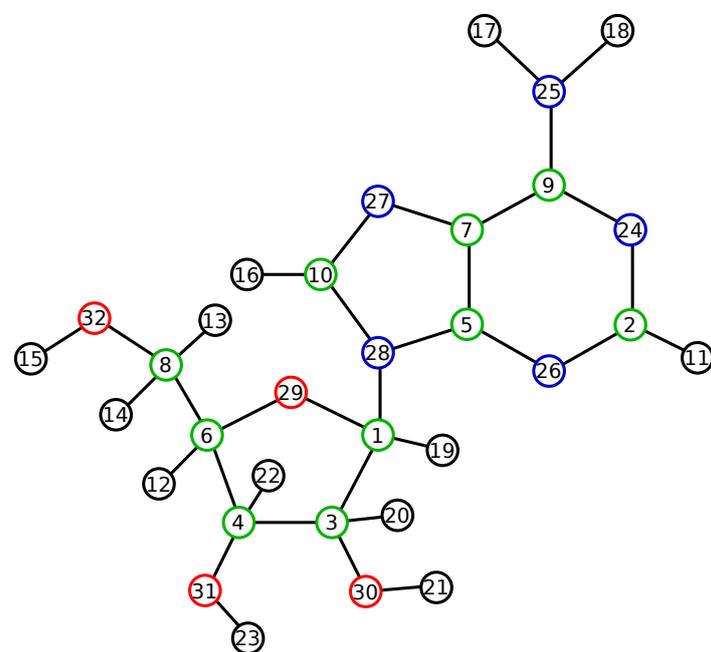
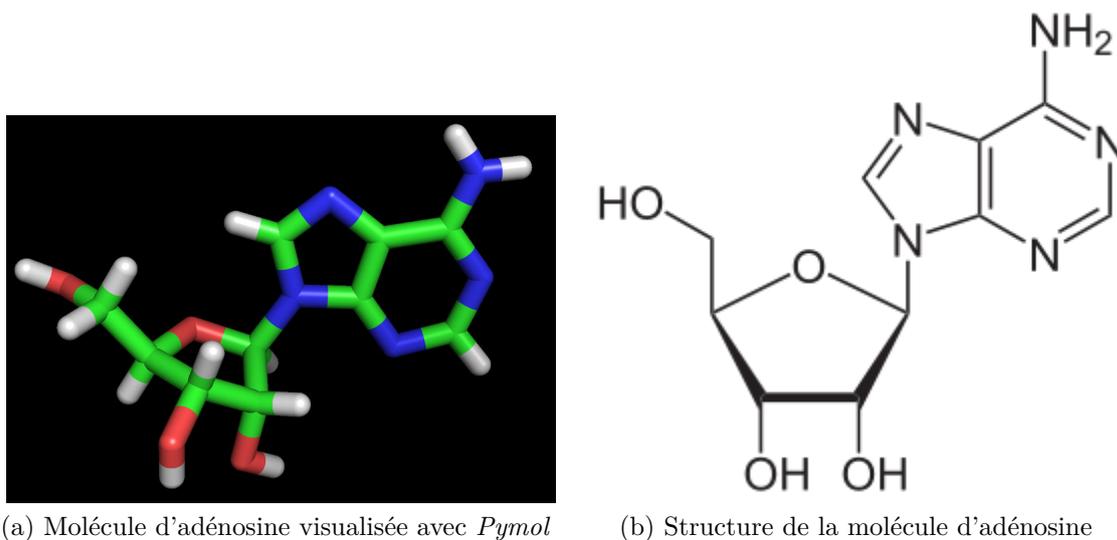
Enfin le dernier module génère la sortie des solutions. Chaque solution est sauvegardée dans un fichier d'extension « mol2 ». Puisqu'il s'agit d'une extension courante en chimie, ces solutions peuvent être directement visualisées à l'aide de logiciels tels que *Pymol*.

L'approche a été testée sur un grand nombre de molécules en utilisant l'application conçue. Nous allons maintenant détailler quelques résultats obtenus sur différentes molécules.

## 2 Étude de l'approche sur la molécule l'adénosine

La molécule d'adénosine, que nous pouvons voir sur la Figure 4.1, est une molécule composée de 32 atomes. La partie droite de la molécule, l'adénine, regroupe des atomes d'hydrogène, de carbone et d'azote. La partie de gauche, le ribose, est constituée d'atomes d'hydrogène, de carbone et d'oxygène. Afin de faciliter la compréhension dans la suite, chaque atome est numéroté.

Cette molécule est constituée de trois cycles différents. Parmi ces trois cycles, deux



(c) Schéma de la molécule d'adénosine avec numérotation des sommets

FIGURE 4.1 – Molécule d'adénosine

sont les cycles aromatiques puisque tous les atomes de ces cycles sont situés dans un même plan. De plus, elle comporte également plusieurs ensembles d'atomes pouvant être impliqués dans des liaisons hydrogènes. Dans cette section, nous allons étudier les différentes étapes de l'approche sur ce substrat.

## 2.1 Construction de l'enveloppe

La première phase de l'approche est la construction de l'enveloppe. Comme expliqué dans le chapitre 2, cette construction s'effectue en trois étapes. La première est la recherche de la topologie de chaque sommet du substrat. L'étape suivante est l'expansion du substrat. C'est-à-dire que pour chaque sommet du substrat, nous cherchons les

positions de leurs extensions en fonction de leur topologie car chacune de ces positions donne lieu à un sommet dans l'enveloppe. Pour finir, l'enveloppe est achevée par la construction de ses arêtes.

### 2.1.1 Topologie des sommets du substrat

Le tableau 4.1 récapitule les topologies des sommets du substrat tels que définis dans la section 2. Nous observons qu'à l'exception des sommets d'hydrogène qui ont toujours une topologie de type (1, 1), les sommets de la partie ribose (celle de gauche sur la Figure 4.1) possèdent tous quatre doublets. Les sommets de carbone ont des topologies de types (4, 0) et ceux d'oxygène sont de types (2, 2). Dans cette partie du substrat, les sommets ont donc une géométrie tétraédrique.

A l'opposé, dans la partie adénine (celle de droite sur la Figures 4.1), les sommets, qui ne sont pas des hydrogènes, possèdent tous trois doublets et ont donc des géométries triangulaires. Les sommets de carbone et deux des sommets d'azote ont une topologie de type (3, 0), alors que les trois autres d'azote ont une topologie de type (2, 1) et possèdent donc un doublet non-liant.

TABLE 4.1 – Topologie des atomes de la molécule d'adénosine.

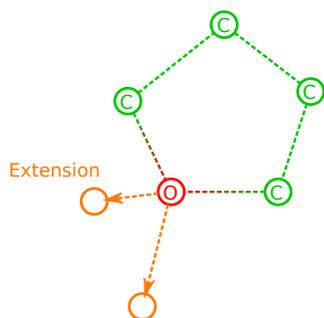
Ribose			Adénine		
n°	Atome	Topologie	n°	Atome	Topologie
1	C	(4, 0)	2	C	(3, 0)
3	C	(4, 0)	5	C	(3, 0)
4	C	(4, 0)	7	C	(3, 0)
6	C	(4, 0)	9	C	(3, 0)
8	C	(4, 0)	10	C	(3, 0)
12	H	(1, 1)	11	H	(1, 1)
13	H	(1, 1)	16	H	(1, 1)
14	H	(1, 1)	17	H	(1, 1)
15	H	(1, 1)	18	H	(1, 1)
19	H	(1, 1)	24	N	(2, 1)
20	H	(1, 1)	25	N	(3, 0)
21	H	(1, 1)	26	N	(2, 1)
22	H	(1, 1)	27	N	(2, 1)
23	H	(1, 1)	28	N	(3, 0)
29	O	(2, 2)			
30	O	(2, 2)			
31	O	(2, 2)			
32	O	(2, 2)			

### 2.1.2 Expansion du substrat

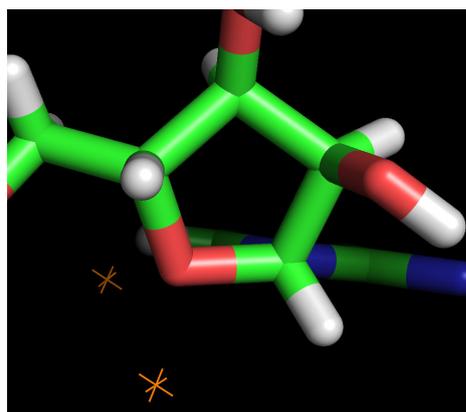
Nous allons maintenant regarder les résultats de la deuxième étape, c'est-à-dire les extensions des sommets du substrat. Pour rappel, les extensions des sommets se situent dans les directions des doublets non-liants des sommets, les seules exceptions étant les

sommets à géométrie triangulaire, c'est-à-dire ceux ayant trois doublets, qui ont également des extensions dans les directions perpendiculaires au plan qu'ils forment avec leurs voisins.

Les sommets de la partie ribose, qui est composée uniquement de sommets d'hydrogène et de sommets à géométrie tétraédrique, ne possèdent que des extensions dans les directions de leurs doublets non-liants. Puisque les sommets de carbone n'ont que des doublets liants, seuls les sommets d'hydrogène et d'oxygène donnent lieu à des extensions. Ces extensions sont représentées sur la Figure 4.2. Chaque sommet d'oxygène a deux extensions alors que les sommets d'hydrogène n'en ont chacun qu'une seule.



(a) Schéma des extensions des sommets de la partie ribose.



(b) Représentation 3D des extensions des sommets du cycle de la partie ribose. Les croix orange représentent les extensions.

FIGURE 4.2 – Extensions des sommets du cycle de la partie ribose de la molécule d'adénosine.

Bien que la partie ribose du substrat comporte un cycle, celui-ci n'est pas conservé dans l'enveloppe. À l'inverse en observant les résultats de la partie adénine du substrat, nous pouvons voir que les deux cycles présents sont conservés dans l'expansion. En effet, tous les sommets de ces cycles sont des sommets à géométrie triangulaire. Qu'importe le nombre de doublets non-liants qu'ils possèdent, ils ont des extensions dans les directions perpendiculaires à leur plan.

La Figure 4.3 montre les extensions des sommets du cycle composé de six atomes de la partie adénine. Chacun des sommets possède aux moins deux extensions perpendiculaires au plan formé par les sommets du cycle. Ce cycle du substrat est projeté deux fois dans l'enveloppe. En plus des extensions perpendiculaires, les deux sommets d'azote du cycle (numéroté 24 et 26 sur la Figure 4.1) possèdent également des doublets non-liants. Ils ont chacun une extension dans la direction de leur doublet non-liant qui se situe dans le même plan que le cycle du substrat.

Le deuxième cycle de la partie adénine est le cycle avec cinq sommets. La Figure 4.4 montre les extensions des sommets de ce cycle. Ici encore, tous les sommets ont une géométrie triangulaire. Comme pour le cycle précédent, chaque sommet possède deux extensions dans les directions perpendiculaires au plan formé par les sommets du cycle. Ainsi ce cycle du substrat est également projeté deux fois dans l'expansion. Le sommet

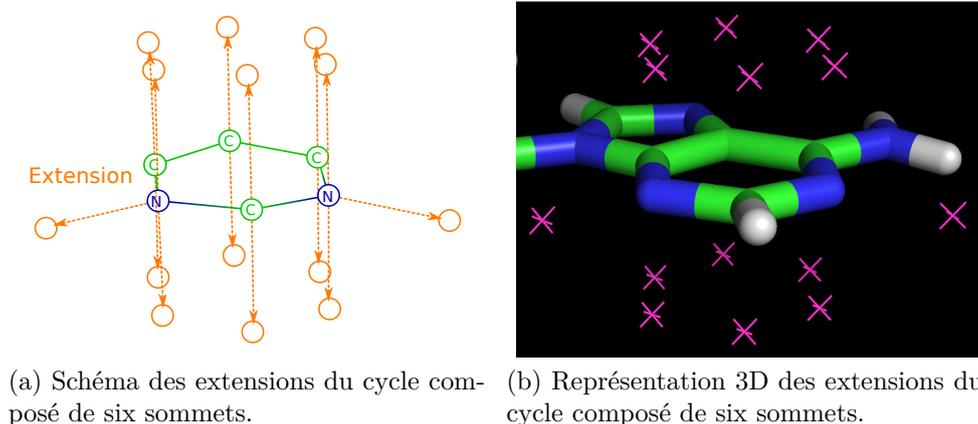


FIGURE 4.3 – Extensions du cycle composé de six sommets de la molécule d'adénosine.

d'azote de topologie (2, 1) (numéroté 25 dans la Figure 4.1) a aussi une extension dans la direction de son doublet non-liant qui est dans le même plan que le cycle du substrat.

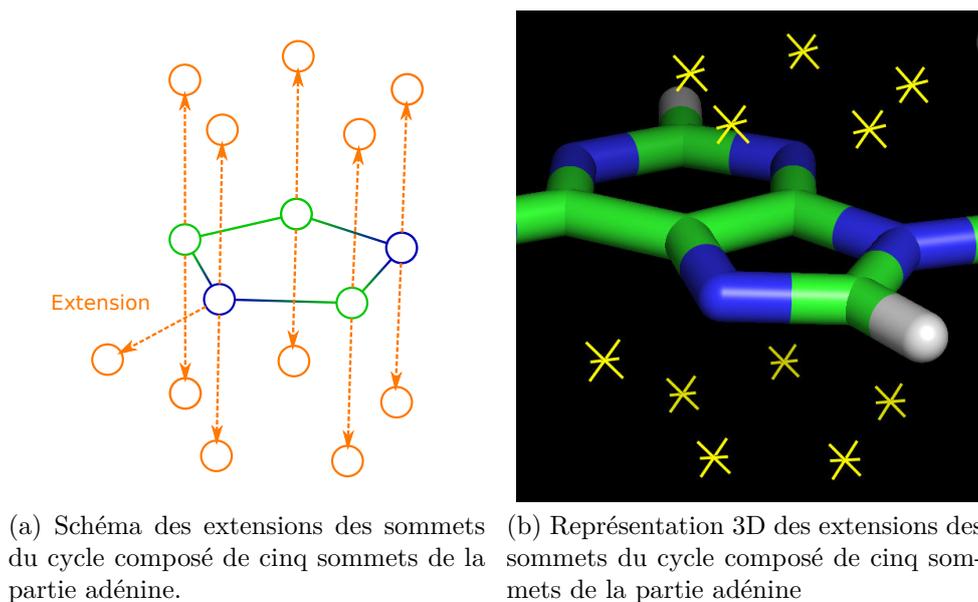


FIGURE 4.4 – Extensions des sommets du cycle composé de cinq sommets de la partie adénine de la molécule d'adénosine.

Au total, 44 extensions sont générées à partir des sommets de la molécule d'adénosine.

### 2.1.3 Construction des arêtes de l'enveloppe

La troisième et dernière étape de cette phase est la construction d'une enveloppe concave à partir des sommets de l'expansion en utilisant la méthode  $\alpha$  - *shape*. Dans les résultats présentés sur la Figure 4.5, le paramètre  $\alpha$  utilisé est 4. Le choix du paramètre  $\alpha$  a une influence sur le nombre de sommets conservés dans l'enveloppe. Plus  $\alpha$  choisi est élevé plus le nombre de sommets restant est faible. Une étude plus poussée

du paramètre  $\alpha$  est réalisée dans annexe B.

Dans cette étape, certains sommets de l'expansion ne sont pas conservés car ils peuvent être situés dans des endroits trop proches du substrat ou simplement inaccessibles par rapport au reste de la molécule.

Dans notre exemple, nous observons ce phénomène sur quelques sommets de l'expansion. Sur la Figure 4.5, six des sommets de l'expansion, n'appartiennent pas à l'enveloppe finale. L'enveloppe finale de la molécule d'adénosine ne comporte donc que 38 sommets. Plus particulièrement, certains des sommets issus des cycles aromatiques du substrat ne sont pas conservés. Nous ne pourrions donc pas obtenir que des cycles aromatiques complets au sein des molécules cages.

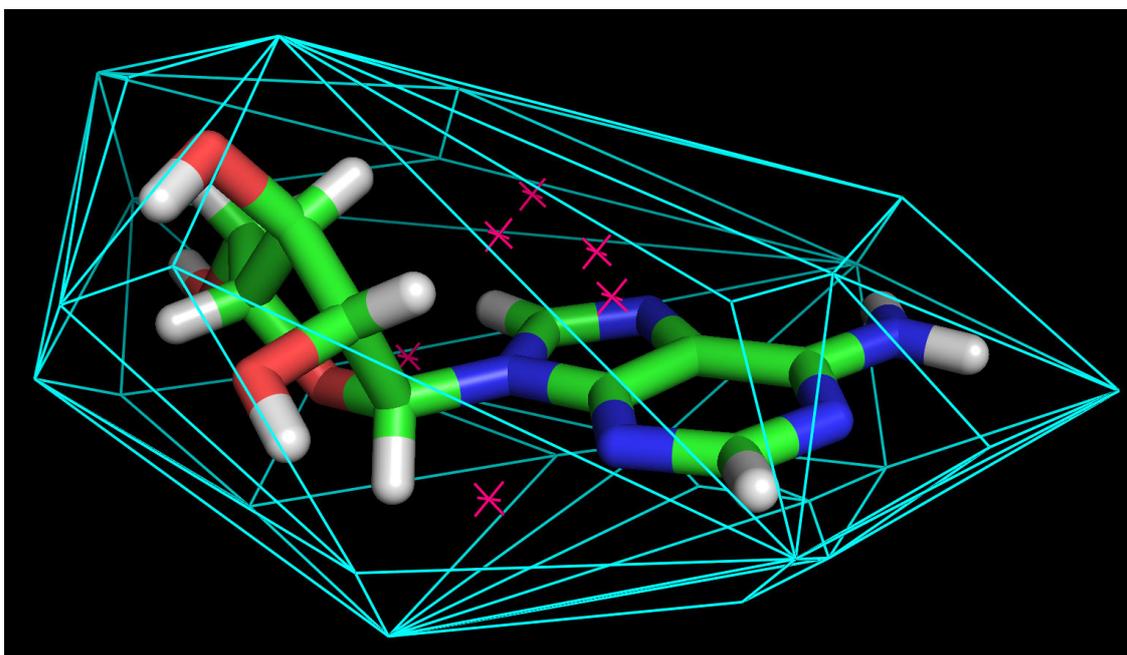


FIGURE 4.5 – Construction des arêtes de l'enveloppe avec  $\alpha = 4$ .

L'enveloppe obtenue est une structure dont la forme est spécifique au substrat. Plus les molécules qui seront générées seront proches de cette structure plus la reconnaissance entre le substrat et les cages sera importante.

## 2.2 Insertion des motifs liants

L'enveloppe est à présent utilisée pour générer des graphes moléculaires qui représenteront les cages moléculaires. La première étape de cette génération est de positionner les motifs liants, c'est-à-dire les motifs moléculaires des graphes qui pourront créer des liaisons avec le substrat. Dans notre approche, deux motifs liants peuvent être insérés. Les motifs aromatiques et les motifs hydrogènes.

### 2.2.1 Insertion des motifs aromatiques

La figure 4.6 montre que sur les 4 cycles qui ont été projetés, seuls deux cycles sont restés complets. Le second cycle projeté à partir de l'anneau à six atomes n'est

que partiel. Et le second cycle projeté à partir du cycle de 5 atomes n'est plus présent. En effet, comme certains sommets des cycles ont été supprimés lors du calcul de l'enveloppe concave les cycles ne peuvent être ajoutés que partiellement dans le graphe final.

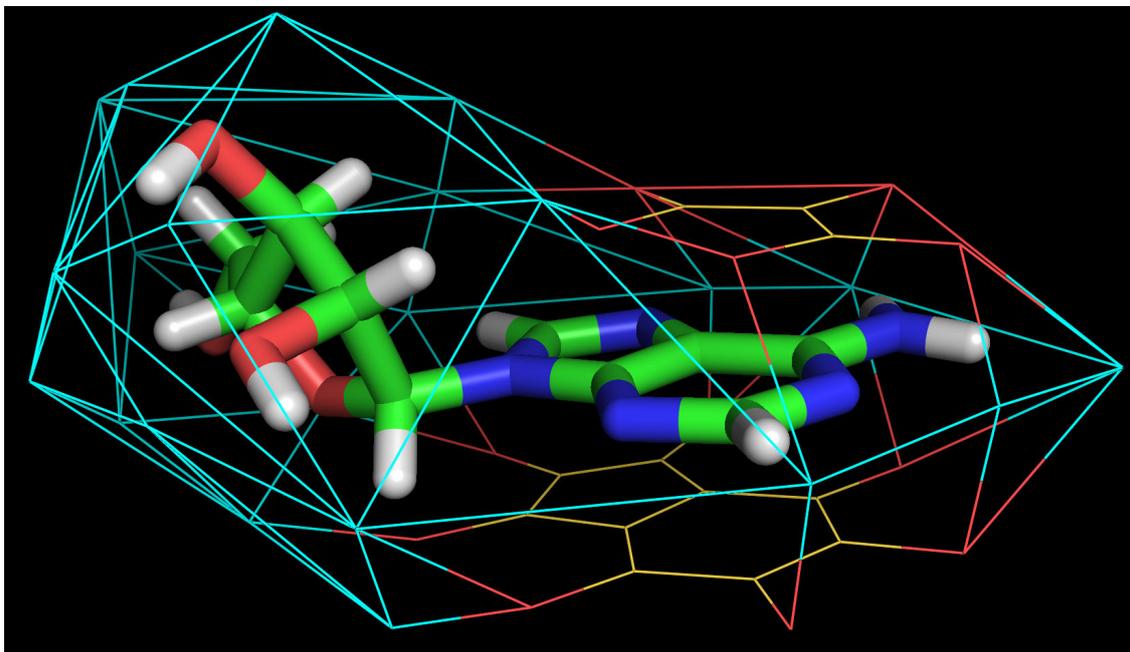


FIGURE 4.6 – Reconstruction des cycles aromatiques de l'enveloppe. Les sommets fixés des motifs sont en orange et les autres sommets des motifs sont en rouge.

Comme nous l'avons expliqué dans la section 2 du chapitre 3, au lieu d'insérer des cycles complets ou partiels dans l'enveloppe, nous insérons des motifs à géométrie triangulaire composés de quatre sommets, un sommet central et ses trois voisins. Ainsi pour construire les cycles aromatiques dans les cages, nous remplaçons les sommets de l'enveloppe étant issus des extensions perpendiculaires des sommets du substrat appartenant à des cycles aromatiques par ces motifs.

### 2.2.2 Insertion des motifs hydrogènes

La première étape de l'insertion des motifs hydrogènes est de trouver quels sont les sommets de l'enveloppe qui doivent être utilisés comme centre pour les insertions de ces motifs. Comme expliqué dans la section 3 du chapitre 3, nous commençons par déterminer quels sont les sommets du substrat qui peuvent participer à une liaison hydrogène en construisant le graphe de dépendance du substrat.

Pour rappel, seuls des atomes d'oxygène, d'azote et de fluor peuvent être accepteurs dans une liaison hydrogène et seuls les atomes d'hydrogène reliés à des hétéroatomes peuvent être donneurs. La Figure 4.7 est une représentation de la molécule d'adénosine, ainsi que de son graphe de dépendance.

La molécule d'adénosine est composée de quatre sommets d'oxygène de topologie (2, 2), ils font donc partie du graphe de dépendance du substrat. Sur les cinq sommets azotes de la molécule, deux d'entre eux ont une topologie (3, 0) et ne possèdent donc

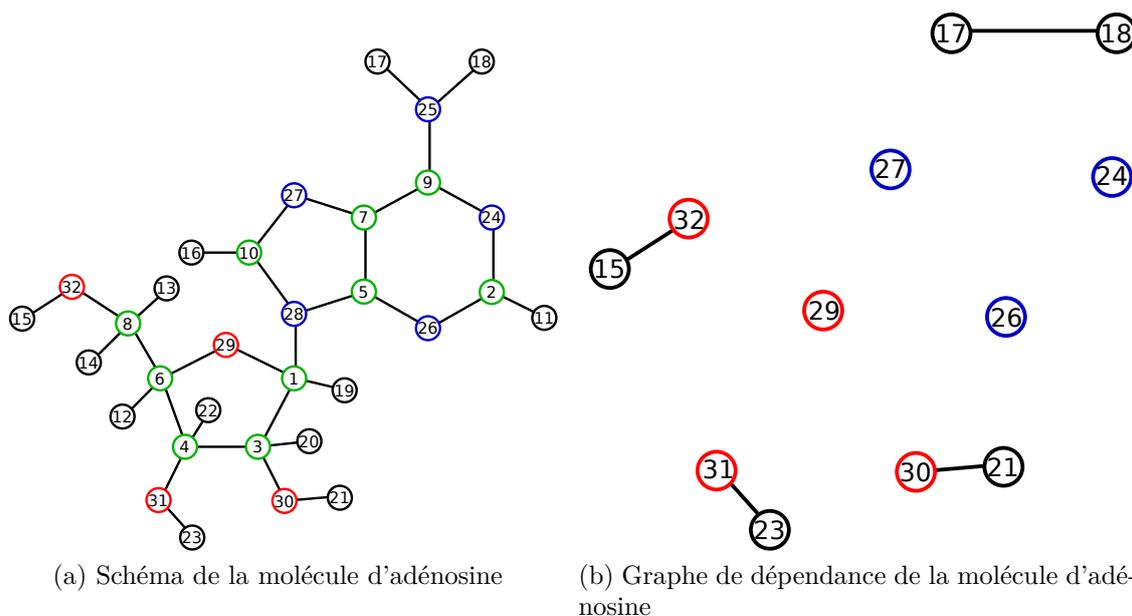


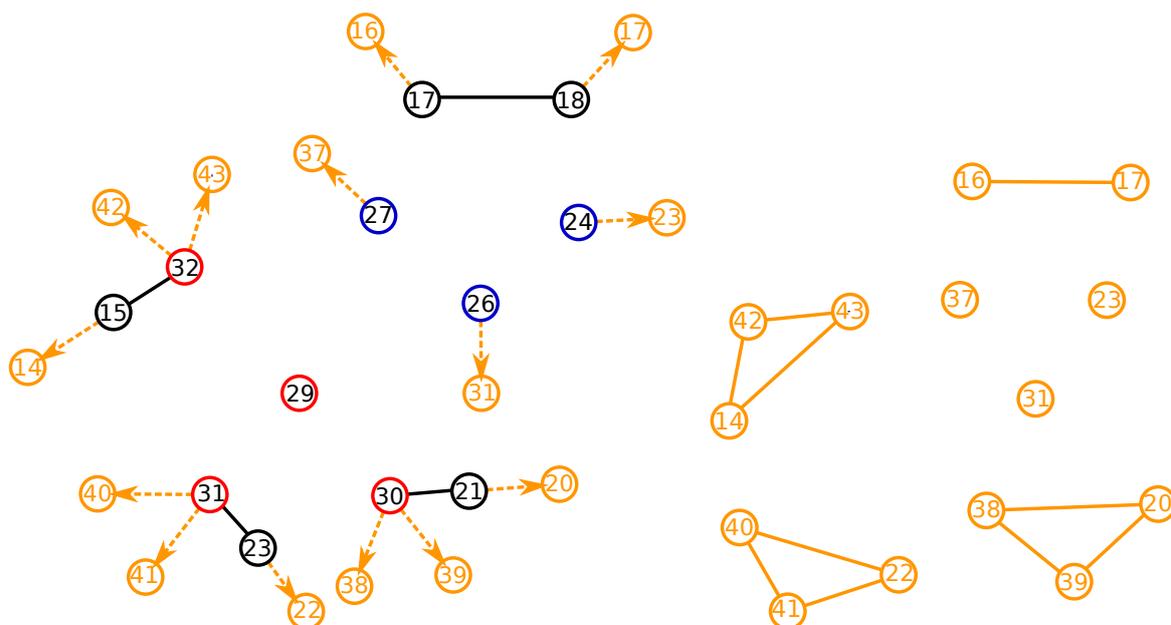
FIGURE 4.7 – Molécule d'adénosine et son graphe de dépendance

pas de doublet non-liants, ils ne sont donc pas dans le graphe de dépendance du substrat. À l'inverse les trois autres sommets azotes ont une topologie  $(2,1)$  et en font donc partie. Au total, sept sommets du graphe sont accepteurs et sont donc intégrés au graphe de dépendance du substrat.

Sur les treize sommets hydrogènes du substrat, seuls cinq d'entre eux ne sont pas reliés à un carbone. Le graphe de dépendance compte donc cinq donneurs. De plus, les sommets hydrogènes 17 et 18 sont reliés à un même sommet d'azote, ils ne peuvent donc pas être utilisés simultanément dans des liaisons hydrogènes. Ils sont alors connectés dans le graphe de dépendance. Il en est de même pour les couples de sommets hydrogène/oxygène. Si le sommet d'oxygène est utilisé comme accepteur, le sommet d'hydrogène ne peut pas être utilisé comme donneur et inversement. C'est pourquoi les arêtes  $[15, 32]$ ,  $[21, 30]$  et  $[23, 31]$  sont présentes dans le graphe de dépendance du substrat.

Le graphe de dépendance du substrat est composé de douze sommets et de quatre arêtes. Chaque sommet d'hydrogène et chaque sommet d'azote donne lieu à une extension. Les sommets d'oxygène, quant à eux, en donnent deux chacun. Cependant les extensions du sommet d'oxygène 29 sont toutes les deux supprimées au moment du calcul de l'enveloppe concave et ne font donc pas partie de l'enveloppe finale. Au total, quatorze sommets font partie du graphe de dépendance de l'enveloppe.

La Figure 4.8a nous montre les extensions des sommets du graphe de dépendance du substrat qui apparaissent dans l'enveloppe. La Figure 4.8b représente le graphe de dépendance de l'enveloppe de la molécule d'adénosine. Si deux extensions sont issues d'un même sommet ou si deux extensions sont issues de deux sommets liés dans le graphe de dépendance du substrat, ils ont une dépendance dans le graphe de dépendance de l'enveloppe. C'est pourquoi des arêtes sont présentes dans le graphe de dépendance de l'enveloppe entre les sommets 16 et 17, ainsi qu'entre les sommets 14,



(a) Graphe de dépendance du substrat avec les extensions (b) Graphe de dépendance de l'enveloppe de ces sommets

FIGURE 4.8 – Graphes de dépendance de la molécule d'adénosine et celui de son enveloppe

42 et 43, les sommets 22, 40 et 41, et enfin les sommets 20, 38 et 39.

Ensuite des copies de l'enveloppe sont réalisées de telle sorte que chaque copie conserve des ensembles de sommets différents dans leur graphe de dépendance. Les sommets 23, 31 et 37 n'ont pas de dépendance et apparaissent donc dans toutes les copies. À l'inverse, la moitié des copies conserveront le sommet 16 et l'autre moitié le sommet 17. Il en est de même pour les cycles de taille trois. Chaque sommet d'un cycle sera conservé dans un tiers des copies. Au total, 54 solutions sont trouvées. Ces solutions donneront donc 54 graphes moléculaires différents.

La Figure 4.9 représente l'une des solutions trouvées après insertion de motifs hydrogènes. Dans cet exemple, les sommets 14, 16, 20, 22, 23, 31 et 37 ont été remplacés par des motifs hydrogènes.

En conclusion, dans cet exemple l'approche a permis de définir une structure englobant le substrat. Contrairement à ce qu'on aurait pu penser, il n'est pas possible d'entourer totalement les cycles aromatiques par d'autres cycles aromatiques identiques. Bien que cela aurait été intéressant pour renforcer les interactions et la spécificité entre les cages moléculaires et le substrat, l'approche montre que cela est géométriquement et chimiquement impossible. De plus, l'approche permet aussi de positionner un total de sept motifs hydrogènes autour du substrat. Chacun d'entre eux permettant également une augmentation de la spécificité des cages qui en découleront par rapport au substrat.

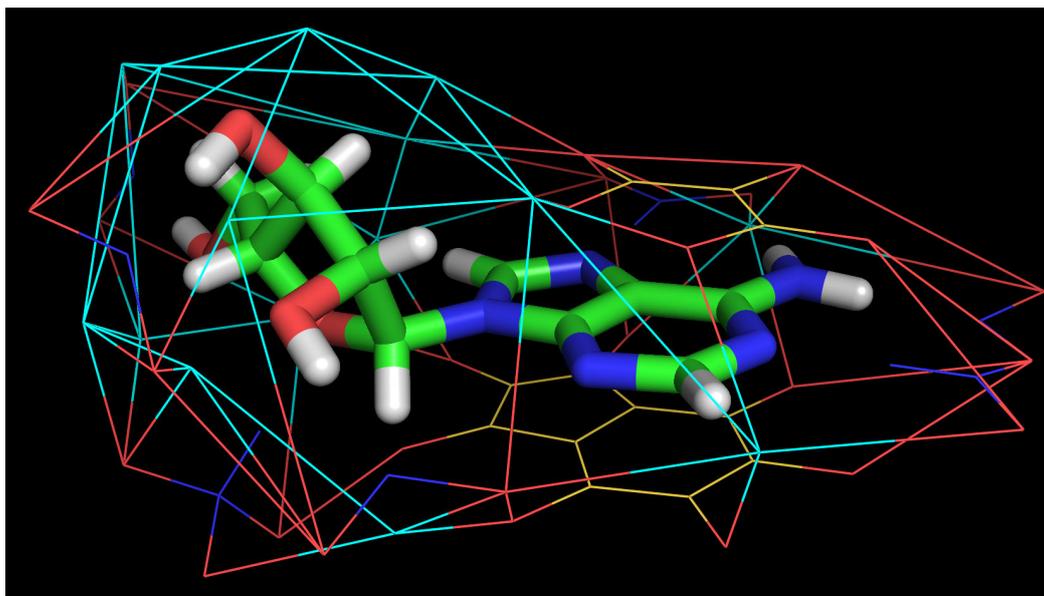


FIGURE 4.9 – Exemple d’une enveloppe après insertion de motifs hydrogènes. Les sommets fixés sont en bleu et les autres sommets des motifs sont en rouge.

### 3 Étude d’approche sur la molécule de saccharose

La molécule de saccharose, que nous pouvons voir sur la Figure 4.10, est un cas de molécule ne comprenant aucun cycle aromatique, mais possédant un grand nombre de sites pouvant être impliqués dans des liaisons hydrogènes. Elle est composée de 45 atomes. Cette molécule regroupe exclusivement des atomes d’hydrogène, de carbone et d’oxygène. Ici encore les atomes sont numérotés.

#### 3.1 Construction de l’enveloppe

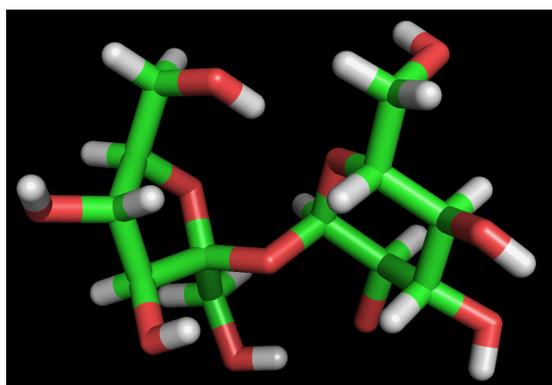
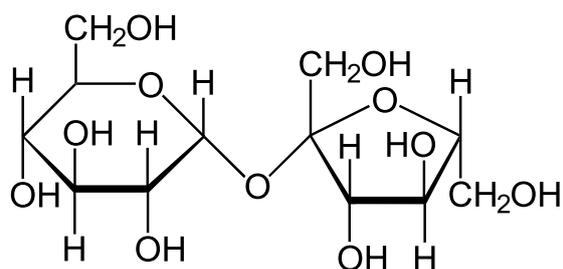
Le tableau 4.2 récapitule les topologies des sommets du substrat tels que définis dans la section 2.

La particularité de la molécule de saccharose est qu’elle est composée exclusivement de sommets à géométrie tétraédrique à l’exception des sommets d’hydrogène. La topologie des sommets est très régulière. Tous les sommets de carbone ont une topologie (4,0) et tous les sommets d’oxygène ont une topologie (2,2).

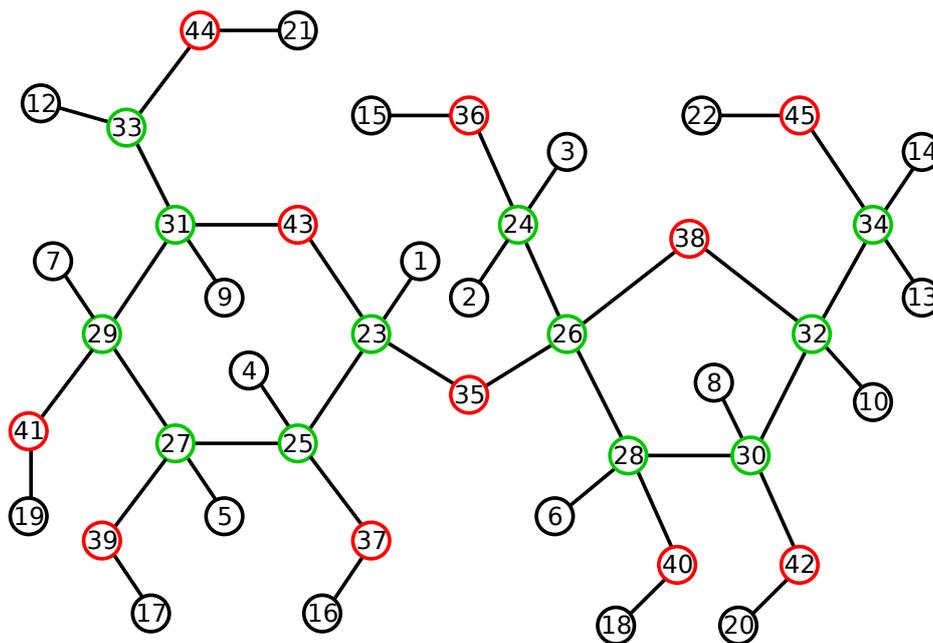
Il n’y a donc aucun cycle aromatique dans la molécule puisque ceux-ci sont composés de sommets à géométrie triangulaire. Les extensions du substrat sont concentrées sur les sommets d’hydrogène et les sommets d’oxygène. La molécule de saccharose compte 44 extensions. Cependant, comme nous le montre la Figure 4.11, sept d’entre elles ne sont pas dans l’enveloppe finale car elles sont trop proches du substrat. Au total, l’enveloppe est composée de 35 sommets.

#### 3.2 Insertion des liaisons hydrogènes

Comme la molécule de saccharose ne possède aucun cycle aromatique, il n’y a pas d’insertion de motifs aromatiques dans les solutions. L’étape suivante est donc l’insertion des motifs hydrogènes.

(a) Molécule de saccharose visualisée avec *Pymol*

(b) Structure de la molécule de saccharose



(c) Schéma de la molécule de saccharose

FIGURE 4.10 – Molécule de saccharose

Comme le montre la Figure 4.12, le graphe de dépendance du substrat est composé de 19 sommets. En effet, parmi les vingt-deux sommets d'hydrogène, seuls huit d'entre eux sont reliés à des sommets d'oxygène (hétéro-atome). En additionnant ces huit sommets d'hydrogène aux onze sommets d'oxygène, on a bien 19 sommets. Chacun des sommets d'hydrogène est relié à un sommet d'oxygène, entraînant huit arêtes dans le graphe.

Étant donné que chacun des sommets d'oxygène donne lieu à deux extensions et que les sommets d'hydrogène en donne une, il y aurait dû y avoir trente sommets dans le graphe de dépendance de l'enveloppe. Cependant sept d'entre eux ont été supprimés lors de la troisième étape de la construction de l'enveloppe. Il reste donc vingt trois sommets dans le graphe de dépendance de l'enveloppe comme nous le montre la Figure 4.13.

Puisque le graphe de dépendance de l'enveloppe possède 5 cliques de taille 3, 3 cliques de taille 2 et 2 cliques de taille 1, il y a donc  $3^5 * 2^3 * 1^2 = 1944$  solutions qui seront générées. Puisqu'il y a dix cliques au total, chaque solution sera composée

TABLE 4.2 – Topologie des atomes de la molécule de saccharose.

n°	Atome	Topo									
1	H	(1, 1)	12	H	(1, 1)	23	C	(4, 0)	35	C	(2, 2)
2	H	(1, 1)	13	H	(1, 1)	24	C	(4, 0)	36	C	(2, 2)
3	H	(1, 1)	14	H	(1, 1)	25	C	(4, 0)	37	C	(2, 2)
4	H	(1, 1)	15	H	(1, 1)	26	C	(4, 0)	38	C	(2, 2)
5	H	(1, 1)	16	H	(1, 1)	27	C	(4, 0)	39	C	(2, 2)
6	H	(1, 1)	17	H	(1, 1)	28	C	(4, 0)	40	C	(2, 2)
7	H	(1, 1)	18	H	(1, 1)	29	C	(4, 0)	41	O	(2, 2)
8	H	(1, 1)	19	H	(1, 1)	30	C	(4, 0)	42	O	(2, 2)
9	H	(1, 1)	20	H	(1, 1)	31	C	(4, 0)	43	O	(2, 2)
10	H	(1, 1)	21	H	(1, 1)	32	C	(4, 0)	44	O	(2, 2)
11	H	(1, 1)	22	H	(1, 1)	33	C	(4, 0)	45	O	(2, 2)
						34	C	(4, 0)			

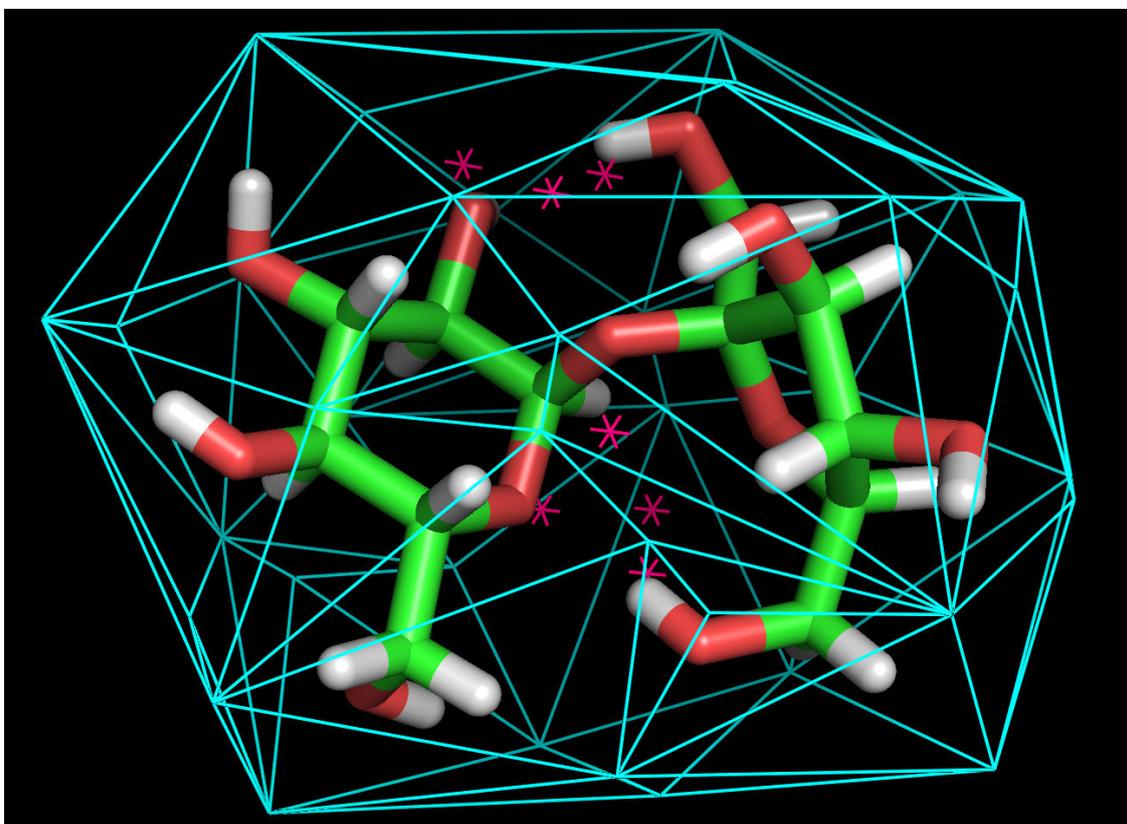


FIGURE 4.11 – Enveloppe de la molécule de saccharose.

de dix sommets. La Figure 4.14 montre l'enveloppe obtenue après insertion de motifs hydrogènes sur les sommets 16, 17, 18, 19, 20, 21, 25, 30, 39 et 44. Cette solution a la particularité de conserver tous les sommets issus de sommets d'hydrogène.

Sur cet exemple, nous pouvons voir que l'approche peut générer un grand nombre de solutions différentes en fonction du nombre de motifs hydrogènes présents dans le substrat de départ.

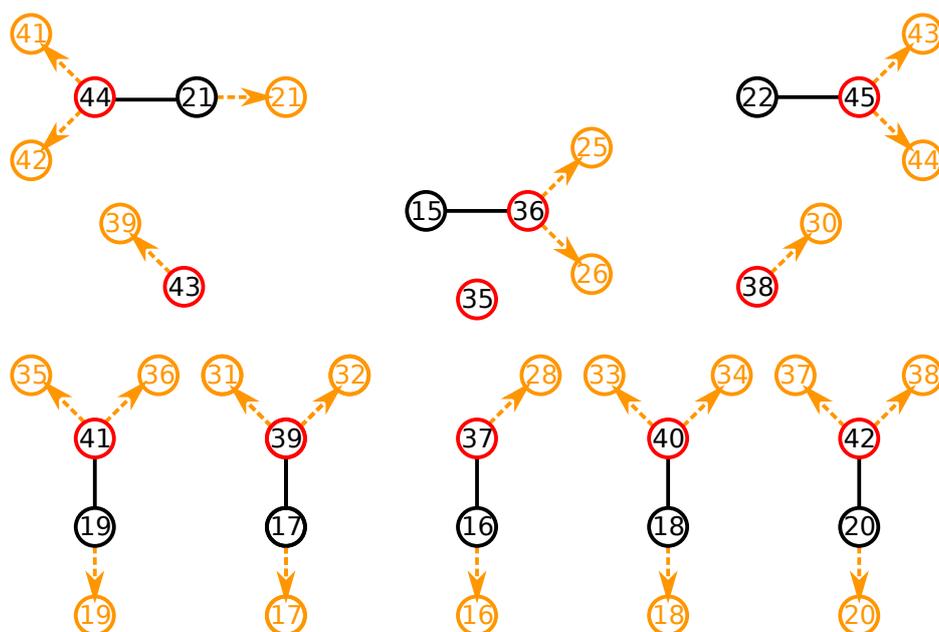


FIGURE 4.12 – Graphe de dépendance de la molécule de saccharose et les extensions conservées dans l’enveloppe de ces sommets.

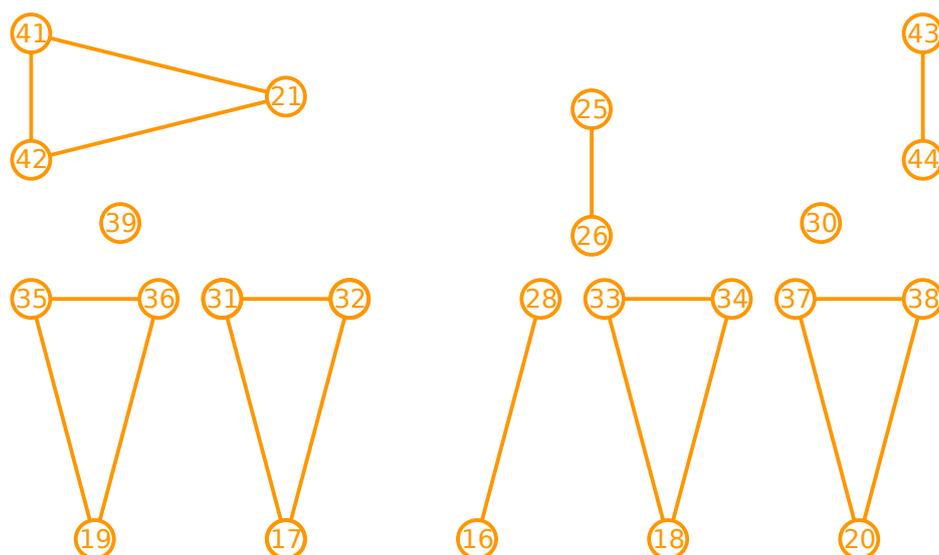


FIGURE 4.13 – Graphe de dépendance de l’enveloppe de saccharose.

## 4 Étude de l’approche sur les molécules d’acétanilide et D-tyrosine

Nous allons maintenant montrer les résultats obtenus sur deux autres molécules qui ont une constitution assez similaire mais qui ont des géométries différentes. La première est la molécule d’acétanilide. C’est une petite molécule composée de 19 atomes, comprenant un cycle aromatique et deux atomes pouvant participer à des liaisons hydrogènes. La seconde molécule est l’acide aminé D-tyrosine. Comme l’acétanilide, le D-tyrosine est composé d’un cycle aromatique et de quelques atomes pouvant participer à des liaisons hydrogènes, pour un total de 24 atomes. Cependant sa forme recourbée nous montre d’autres intérêts de l’approche.

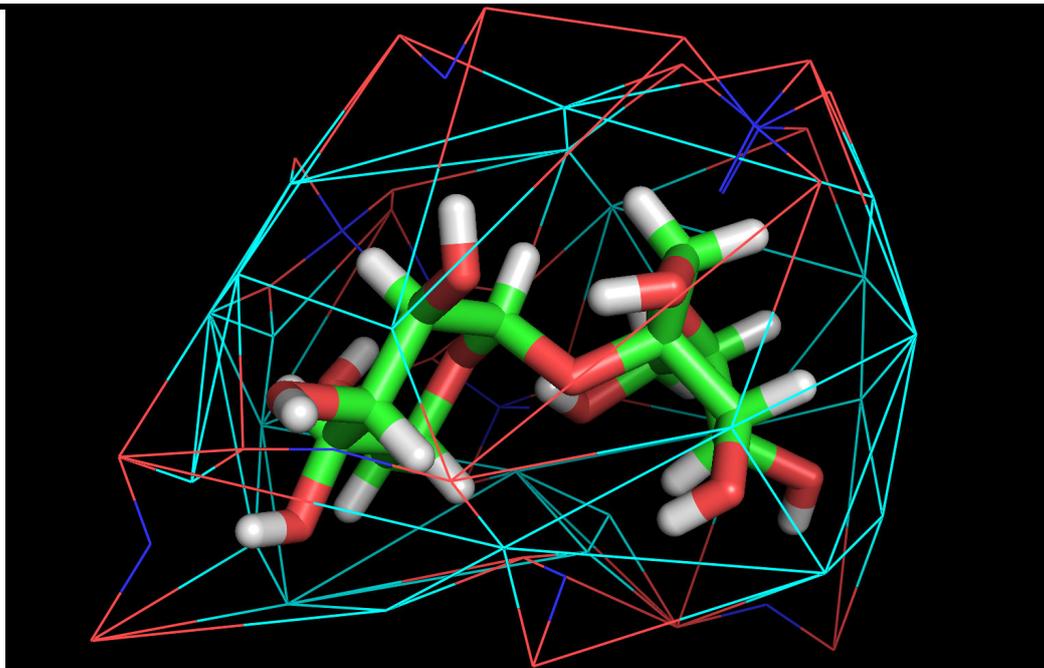


FIGURE 4.14 – Exemple d’une enveloppe après insertion de motifs hydrogènes.

#### 4.1 Exemples de la molécule d’acétanilide

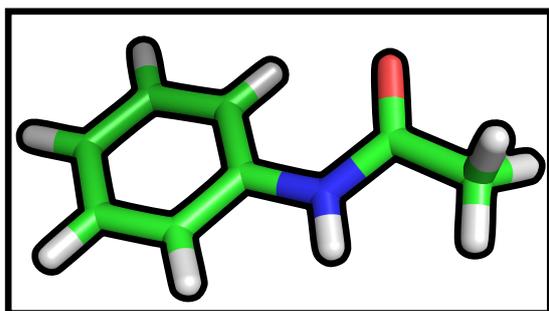
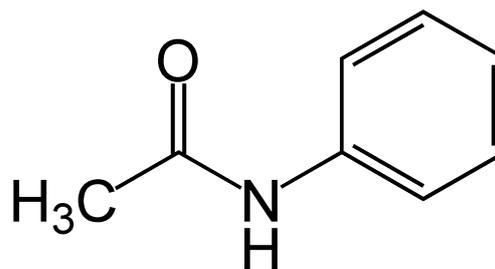
L’acétanilide, que nous pouvons voir sur la Figure 4.15, est une molécule très utilisée en pharmaceutique. Cette molécule est composée de 19 atomes dont huit carbones, neuf hydrogènes, un azote et un oxygène. Parmi les carbones, six d’entre eux forment un cycle aromatique. De plus certains atomes de la molécule peuvent être impliqués dans des liaisons hydrogènes. Dans cette exemple, les deux types de motifs liants peuvent être ajoutés.

Le tableau 4.3 récapitule la topologie des sommets du substrat. À l’exception des hydrogènes qui sont toujours de topologie (1, 1) et d’un carbone de topologie (4, 0), tous les autres possèdent trois doublets. Les sommets représentant les carbones ainsi que celui représentant l’azote sont de topologie (3, 0) alors que le sommet représentant l’oxygène est de topologie (1, 2).

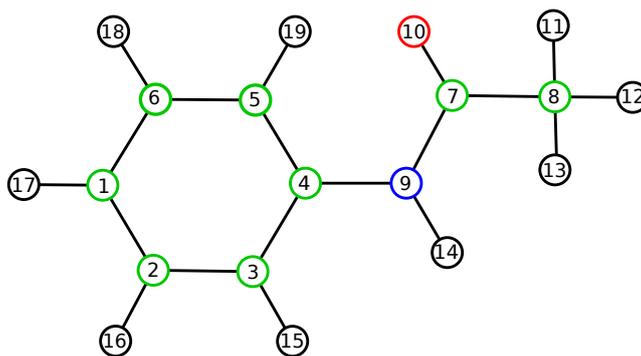
TABLE 4.3 – Topologie des atomes de la molécules d’acétanilide.

n°	Atome	Topologie	n°	Atome	Topologie
1	C	(3, 0)	11	H	(1, 1)
2	C	(3, 0)	12	H	(1, 1)
3	C	(3, 0)	13	H	(1, 1)
4	C	(3, 0)	14	H	(1, 1)
5	C	(3, 0)	15	H	(1, 1)
6	C	(3, 0)	16	H	(1, 1)
7	C	(3, 0)	17	H	(1, 1)
8	C	(4, 0)	18	H	(1, 1)
9	N	(3, 0)	19	H	(1, 1)
10	O	(1, 2)			

L’expansion du substrat donne un total de 29 extensions. Neuf sont issues des som-

(a) Molécule d'acétanilide visualisée avec *Pymol*

(b) Structure de la molécule d'acétanilide



(c) Schéma de la molécule d'acétanilide

FIGURE 4.15 – Molécule d'acétanilide

ments hydrogènes, et deux sont issues des doublets non liants de l'oxygène. Le reste des sommets de l'enveloppe provient des extensions perpendiculaires aux sommets à géométrie triangulaires. Comme le montre la Figure 4.16, les extensions des carbones du cycle aromatique, soit les extensions des sommets 1 à 6, forment deux cycles dans l'enveloppe qui sont des translations du cycle de départ du substrat.

De plus, la figure montre aussi qu'aucun sommet de l'expansion n'est supprimé au moment du calcul de l'enveloppe concave. Le nombre de sommets de l'expansion est donc le même que celui de l'enveloppe finale.

L'étape suivante est la reconstruction des cycles aromatiques de l'enveloppe. Comme dit précédemment aucune des extensions n'est supprimée de l'enveloppe. Puisque le substrat possède un cycle aromatique, cela signifie que deux cycles aromatiques sont également présents dans l'enveloppe. La Figure 4.17 montre que ces deux cycles sont bien retrouvés dans l'enveloppe de part et d'autre de celui du substrat.

Enfin, vient l'insertion des motifs hydrogènes. Parmi les sommets du substrat, seuls deux d'entre eux peuvent être utilisés pour des liaisons hydrogènes. Il s'agit de l'oxygène et de l'hydrogène numéro 14 qui est relié à l'azote. L'azote quant à lui ne possède pas de doublets non-liants, il ne peut donc pas être utilisé comme donneur dans une liaison hydrogène. Le graphe de dépendance du substrat est donc composé de seulement deux sommets qui ne sont pas liés entre eux.

Comme le montre la Figure 4.18, les deux sommets du graphe de dépendance du substrat donnent lieu à trois sommets dans le graphe de dépendance de l'enveloppe. Puisque deux d'entre eux sont issus de l'oxygène, ils sont liés dans le graphe de dé-

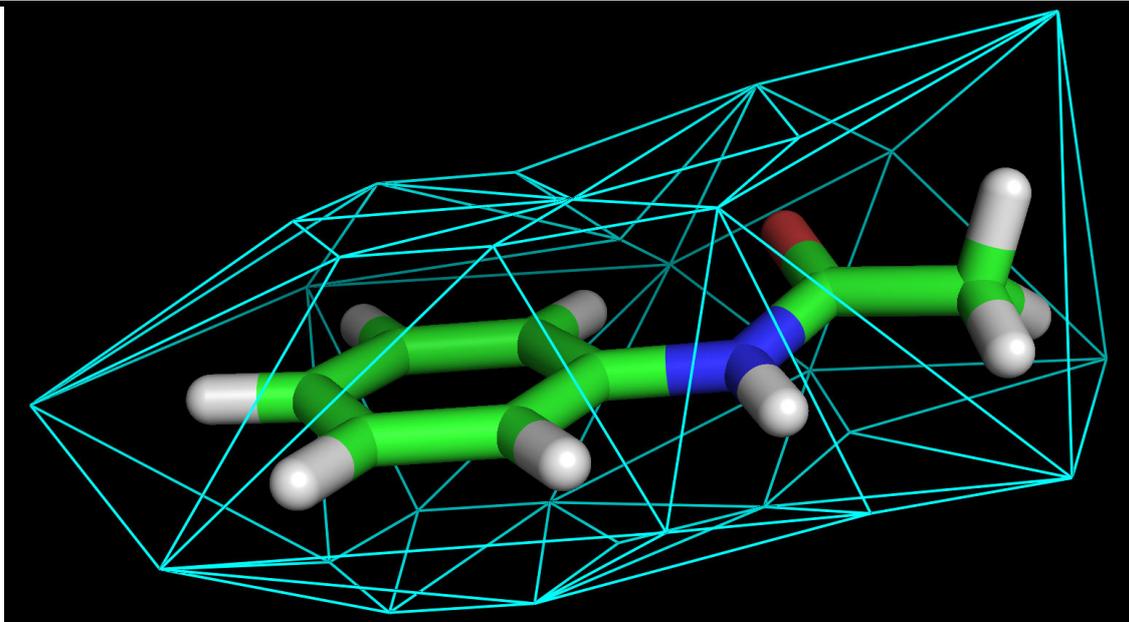


FIGURE 4.16 – Enveloppe de l'acétanilide

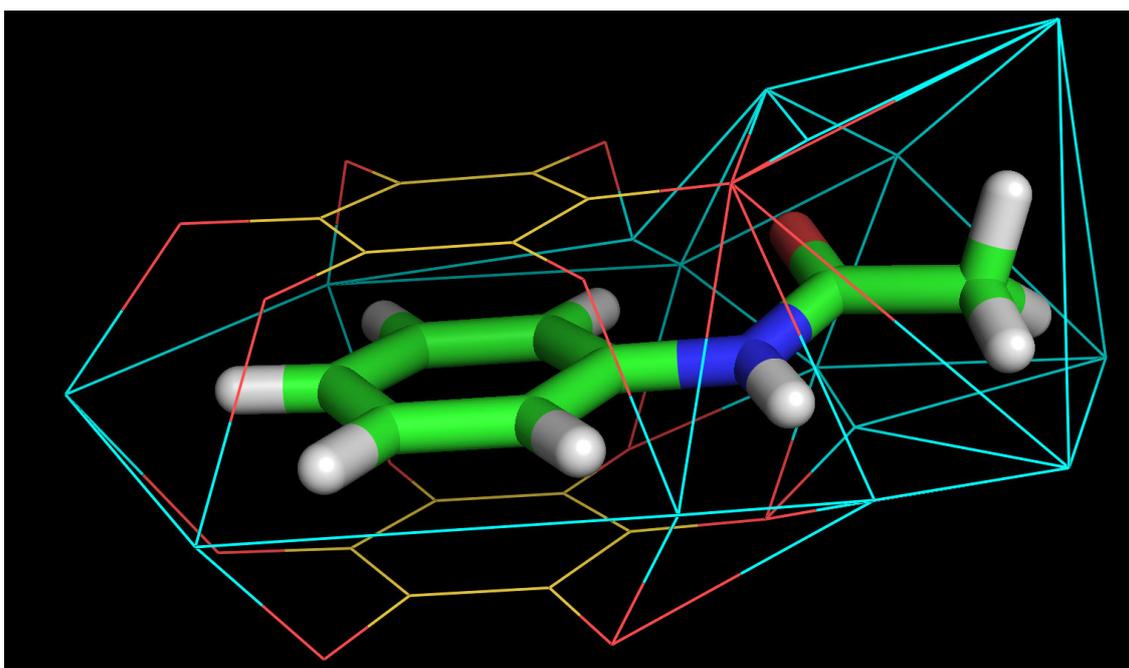


FIGURE 4.17 – Reconstruction des cycles aromatiques de l'enveloppe.

pendance de l'enveloppe.

On obtient donc deux solutions. la première a les sommets 16 et 20 comme sommets pour l'insertion des motifs hydrogènes. Et la deuxième a 17 et 20.

Cependant, dans le cas de la première solution, certains sommets du motif inséré sur le sommet 16 se retrouvent à l'intérieur de la zone délimitée par l'enveloppe. Cette solution est donc éliminée puisqu'elle n'est pas réaliste. La Figure 4.19 montre les résultats de la seconde solution.

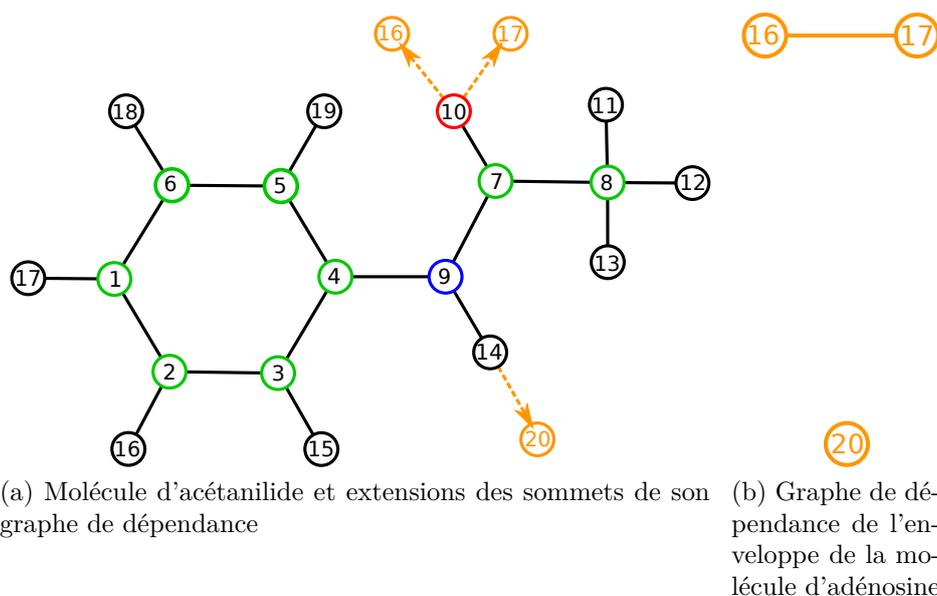


FIGURE 4.18 – Extensions des sommets du graphe de dépendance de la molécule d'acétanilide et graphe de dépendance de son enveloppe

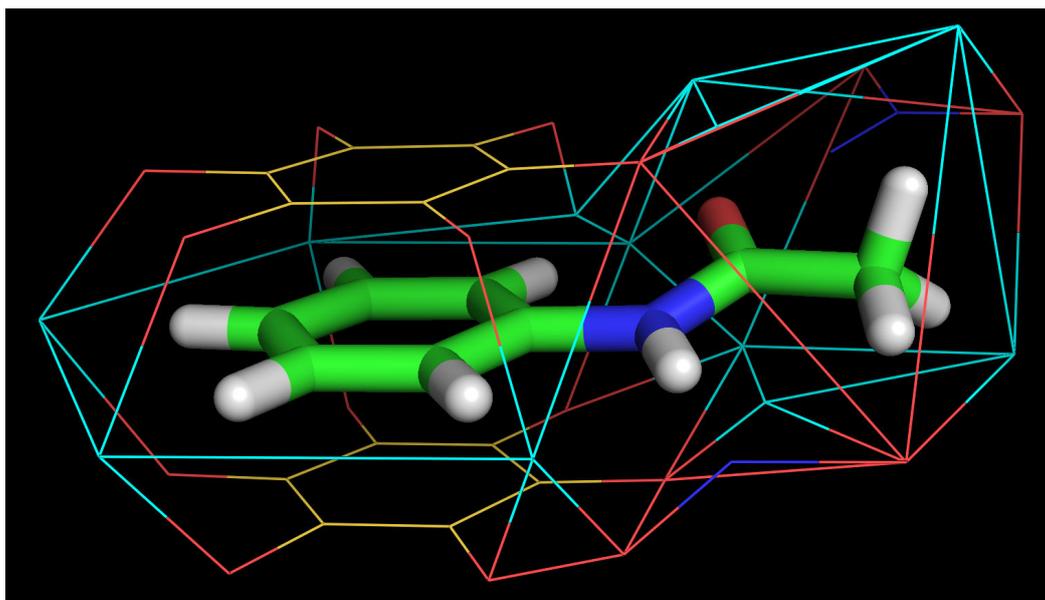
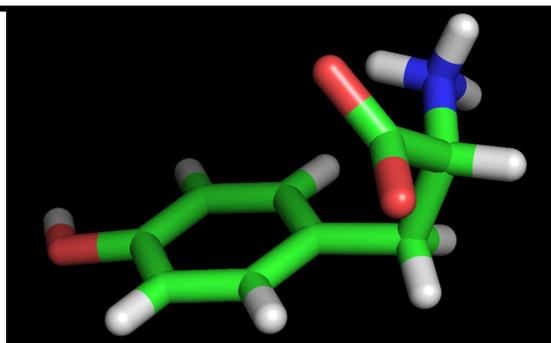
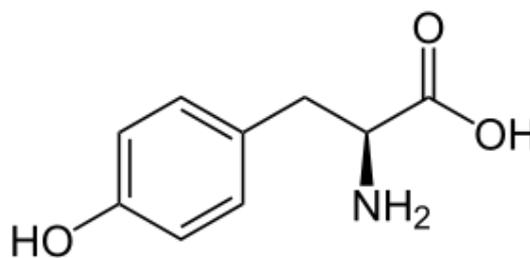


FIGURE 4.19 – Exemple d'une enveloppe après insertion de motifs hydrogènes

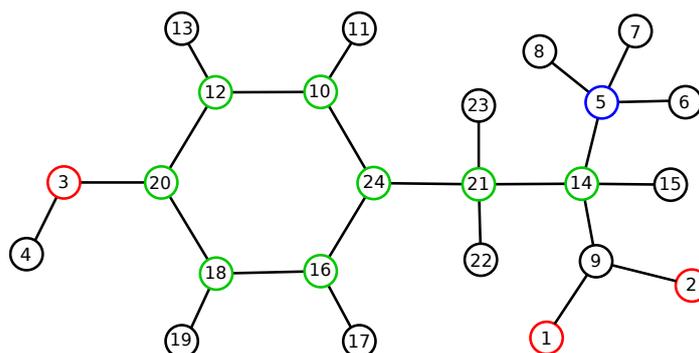
## 4.2 Exemples de la molécule de D-tyrosine

La D-tyrosine est un acide aminé (voir Figure 4.20). Composée de 24 atomes, elle possède neuf atomes de carbone, onze atomes d'hydrogène, trois atomes d'oxygène et un d'azote. Pourvue d'un cycle aromatique et de plusieurs atomes pouvant participer à des liaisons hydrogènes, les solutions de cette molécule peuvent avoir les deux types de motifs liants.

Le tableau 4.4 récapitule la topologie des sommets du substrat. La D-tyrosine possède sept sommets de carbone ayant trois doublets et six d'entre eux forment un cycle aromatique. Les deux carbones restants ainsi que l'azote sont de topologie (4, 0) et

(a) Molécule de D-tyrosine visualisée avec *Pymol*

(b) Structure de la molécule de D-tyrosine



(c) Schéma de la molécule de D-tyrosine

FIGURE 4.20 – Molécule de D-tyrosine

n'ont donc pas d'extension. Enfin, tous les oxygènes possèdent deux doublets non-liants : deux sont de topologie (1,2) et ont donc une géométrie triangulaire alors que le dernier a une géométrie tétraédrique et une topologie (2,2).

L'expansion du substrat donne un total de 35 extensions. Cependant sur ces 35 extensions, 7 n'apparaissent pas dans l'enveloppe finale. En effet la forme courbée de la molécule de D-tyrosine fait que plusieurs des extensions se retrouvent dans une zone non accessible par d'autres atomes. Ces extensions ne font donc pas partie de l'enveloppe concave calculée. Ils sont alors retirés de l'enveloppe. L'enveloppe comporte seulement 29 sommets.

Puisque la D-tyrosine possède un cycle aromatique, les extensions des sommets de ce cycle ont formé deux cycles lors de l'expansion. Cependant, plusieurs des sommets de l'un de ces cycles ne font plus partie de l'enveloppe. Comme le montre la Figure 4.22 l'un des deux cycles est conservé et reconstruit dans l'enveloppe alors qu'un seul des sommets du deuxième cycle apparaît.

Le graphe de dépendance de la D-tyrosine, Figure 4.23, montre que sept sommets peuvent donner lieu à des liaisons hydrogènes. On retrouve les trois oxygènes qui peuvent être accepteurs (l'azote ne le pouvant pas car il ne possède pas de doublet non-liant) et quatre hydrogènes, les trois liés à l'azote et celui lié à l'oxygène 3. Dans ce graphe l'oxygène 3 et l'hydrogène 4 sont liés. Il en est de même pour les hydrogènes 6, 7 et 8 qui sont reliés à l'azote.

Ces sommets du graphe de dépendance donnent dix extensions dans l'enveloppe

TABLE 4.4 – Topologie des atomes de la molécules de D-tyrosine.

n°	Atome	Topologie	n°	Atome	Topologie
1	O	(1,2)	13	H	(1,1)
2	O	(1,2)	14	C	(4,0)
3	O	(2,2)	15	H	(1,1)
4	H	(1,1)	16	C	(3,0)
5	N	(4,0)	17	H	(1,1)
6	H	(1,1)	18	C	(3,0)
7	H	(1,1)	19	H	(1,1)
8	H	(1,1)	20	C	(3,0)
9	C	(3,0)	21	C	(4,0)
10	C	(3,0)	22	H	(1,1)
11	H	(1,1)	23	H	(1,1)
12	C	(3,0)	24	C	(3,0)

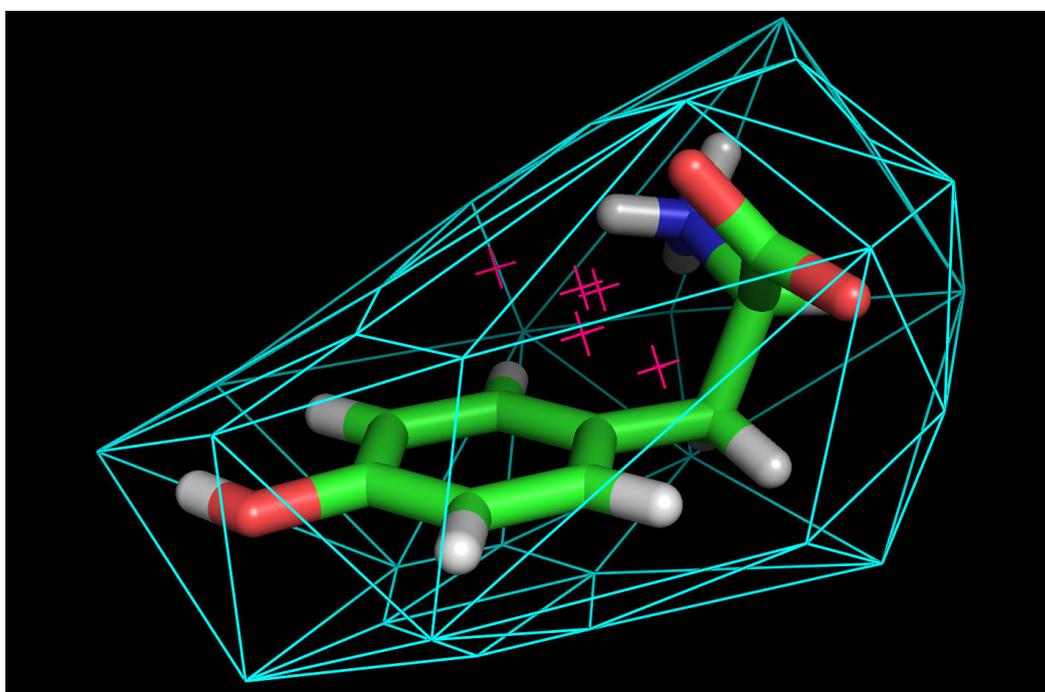


FIGURE 4.21 – Enveloppe de la molécule de D-tyrosine

qui apparaissent dans le graphe de dépendance de celle-ci (voir Figure 4.24). L'oxygène 1 donne deux extensions qui sont dépendantes. De même pour l'oxygène 2. L'oxygène 3 donne aussi deux extensions mais elles sont également reliées à l'extension de l'hydrogène 4 puisque les sommets 3 et 4 du substrat sont dépendants. Enfin les trois sommets issus des trois hydrogènes 6, 7 et 8 sont également reliés dans le graphe de dépendance de l'enveloppe.

Au total 36 solutions peuvent être trouvées pour la molécule de D-tyrosine. La Figure 4.25 montre l'une de ces solutions. Dans cette solution les sommets 1, 5, 9 et 13 sont remplacés par des motifs hydrogènes.

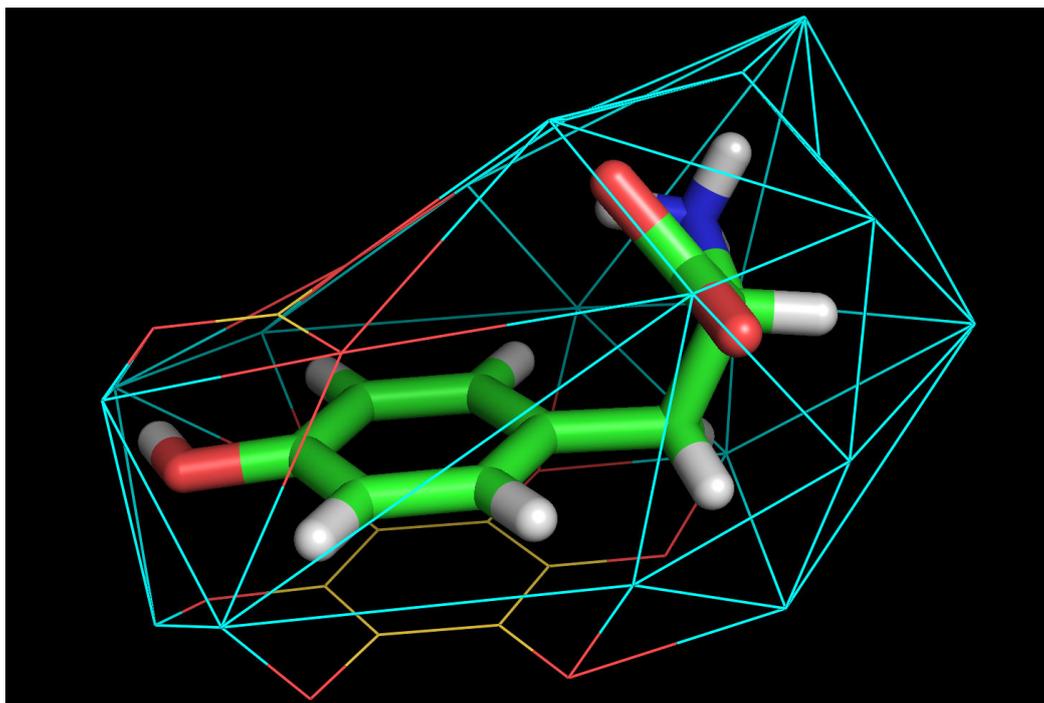


FIGURE 4.22 – Reconstruction des cycles aromatiques de l’enveloppe.

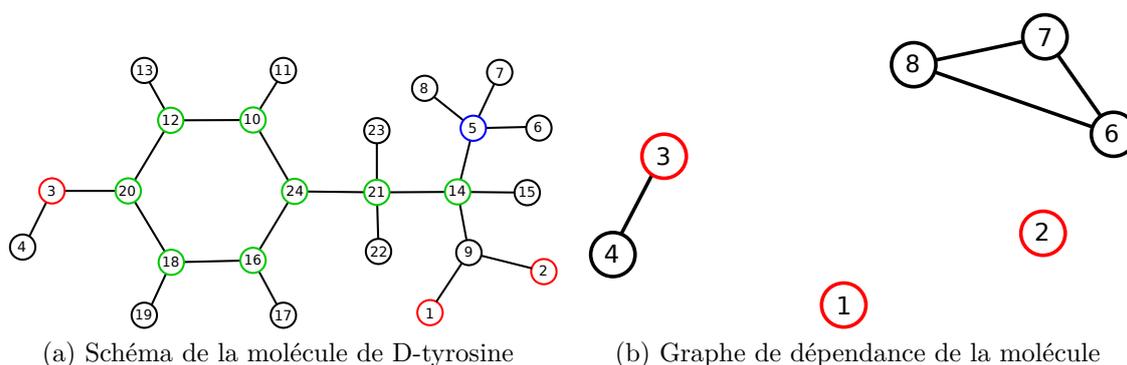


FIGURE 4.23 – Molécule de D-tyrosine et son graphe de dépendance

## 5 Conclusion

Les études de ce chapitre montrent que l’approche que nous avons développé permet de construire une enveloppe autour du substrat qui peut servir de base pour la génération des cages moléculaires. Cette enveloppe est construite de telle manière que ces sommets peuvent être utilisés pour positionner des motifs moléculaires capables d’interagir avec le substrat. Cette enveloppe délimite également une zone autour du substrat à l’intérieur de laquelle les cages doivent être construites afin de garder une cohérence chimique.

Les motifs aromatiques ne peuvent être construits que dans la mesure où le substrat possède au moins un cycle aromatique. Cependant il arrive fréquemment que les sommets de ces cycles ne soient pas conservés dans l’enveloppe. Bien que les interactions soient plus fortes avec des cycles complets, le fait de les construire partiellement

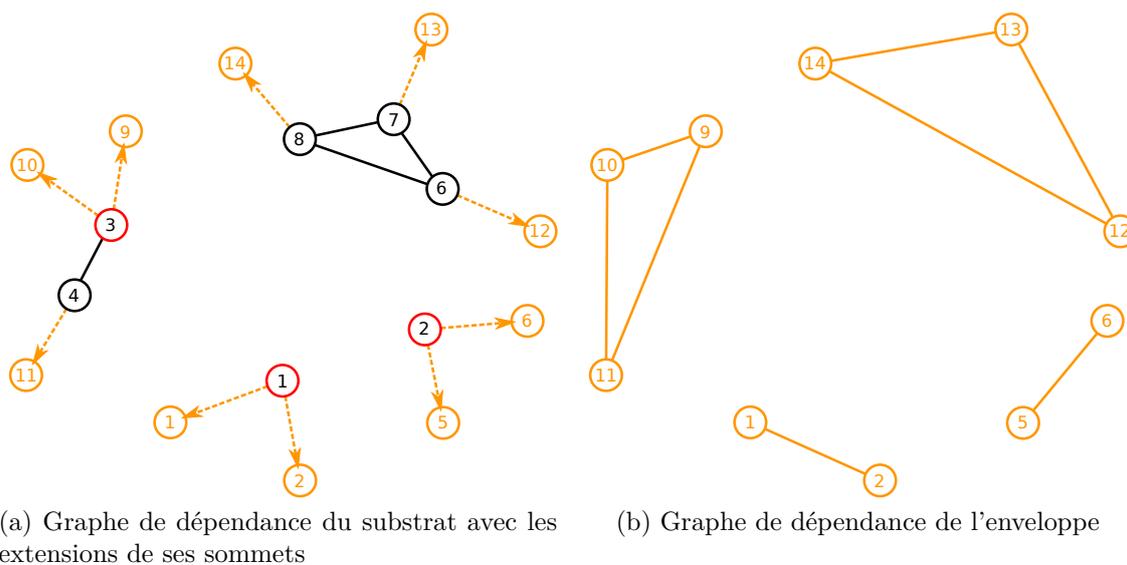


FIGURE 4.24 – Molécule de D-tyrosine et graphe de dépendance de son enveloppe

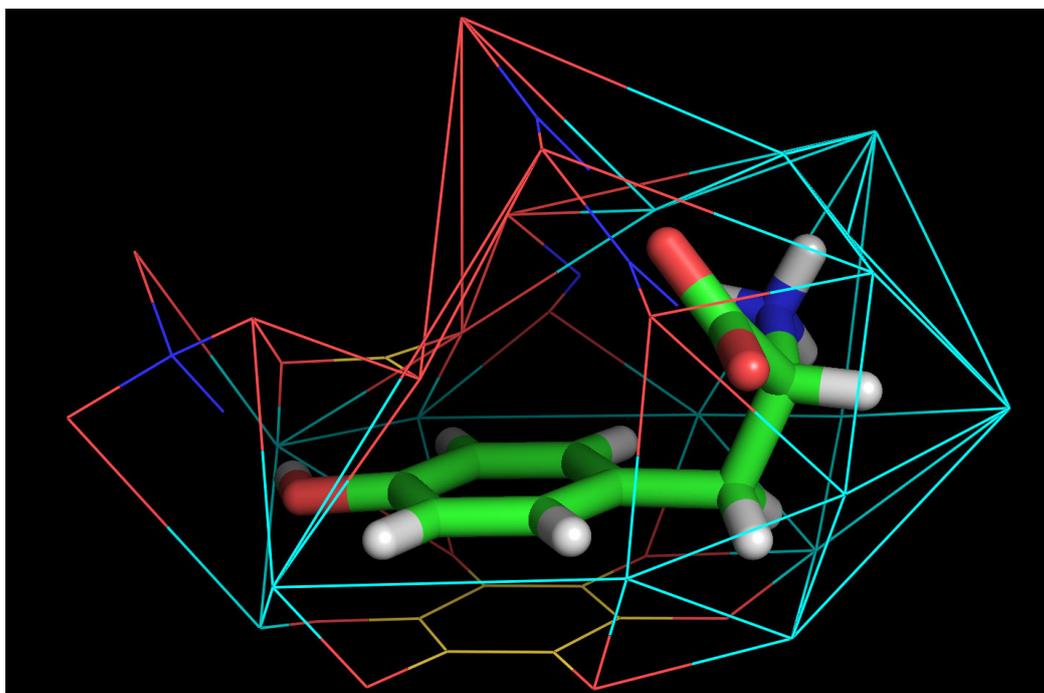


FIGURE 4.25 – Exemple d'une enveloppe après insertion de motifs hydrogènes

permet de pouvoir fixer en partie la forme des molécules cages générées pour qu'elles soient le plus proche possible de celle du substrat tout en favorisant les interactions potentielles avec le cycle.

L'insertion des motifs hydrogènes a un lien direct avec le nombre de solutions générées. Plus le substrat possède un grand nombre de sommets pouvant être impliqués dans des liaisons hydrogènes et plus ces sommets sont dépendants entre eux, plus le nombre de solutions générées augmente. Le nombre de solutions est dépendant du nombre et de la taille des cycles du graphe de dépendance de l'enveloppe puisque le nombre de solutions est le produit de la taille des cycles. Cependant, si lors d'une insertion les

---

sommets du motifs se retrouvent à l'intérieur de la zone délimitée par l'enveloppe, la solution est non réalisable d'un point de vue chimique.



# Conclusion

Dans cette thèse, nous nous sommes intéressés aux problèmes de conception de cages moléculaires et plus précisément à la génération de cages moléculaires pour un substrat donné en insistant sur l'aspect spécifique des cages pour le substrat.

## Résumé de l'approche proposée

L'approche que nous proposons permet la génération de guides pour la construction de cages moléculaires pour un substrat donné. La première phase de cette approche construit une structure qui englobe le substrat et dont la forme est très spécifique à celle du substrat. Cette structure, que nous appelons *enveloppe*, peut être considérée comme une « seconde peau » pour le substrat. Cependant, bien que cette enveloppe ne soit pas une molécule, elle est construite de telle sorte qu'elle puisse servir de support pour la génération des cages moléculaires. Elle est donc construite dans une logique moléculaire. Chacun des sommets qui composent l'enveloppe est positionné dans une direction où les atomes (sommets) du substrat peuvent encore interagir avec d'autres atomes, c'est-à-dire dans une des directions des doublets non-liants d'un des sommets du substrat. De plus, chacun des sommets de l'enveloppe est également positionné à une distance équivalente à celle d'une liaison faible des sommets du substrat. La distance choisie est celle d'une liaison hydrogène car c'est la plus fréquente des liaisons faibles.

Pour ce faire, nous commençons, dans une première étape, par chercher ce que nous appelons la *topologie* des sommets du substrat. En fonction de leur nombre de voisins (doublets liants), de l'angle géométrique qui les sépare et du type d'atomes dont il s'agit, nous déterminons leur nombre de doublets non-liants. En connaissant le nombre de doublets total d'un sommet (nombre de doublets liants et non-liants), nous trouvons les directions des doublets non-liants. Dans une seconde étape que nous appelons *expansion* du substrat, nous positionnons dans chacune de ces directions un sommet de l'enveloppe à une distance équivalente à celle d'une liaison hydrogène.

Enfin, dans une troisième étape, nous construisons une enveloppe concave à partir des sommets de l'enveloppe, trouvés dans l'étape précédente, en utilisant la méthode  *$\alpha$ -shape*. Cette étape nous permet de pouvoir délimiter une zone (l'enveloppe concave) à l'intérieur de laquelle les sommets des cages moléculaires ne pourront pas être positionnés car cela serait incohérent chimiquement. De même, les sommets de l'enveloppe qui ne font pas partie de la zone extérieure de l'enveloppe concave sont supprimés de l'enveloppe car ils sont jugés trop proches du substrat ou inatteignables.

La phase de la construction de l'enveloppe se fait donc en trois étapes : la recherche

---

de la topologie des sommets du substrat, l'expansion du substrat et la construction d'une enveloppe concave à partir des sommets de l'expansion. Ensuite, les deux phases suivantes servent à générer les cages moléculaires à partir de cette enveloppe en assemblant des petits modules moléculaires actifs.

La deuxième phase consiste à positionner des petits motifs moléculaires actifs *liants* autour du substrat en utilisant les sommets de l'enveloppe. Ces motifs sont appelés liants car ce sont à travers eux que les cages moléculaires générées et le substrat pourront interagir en créant des liaisons faibles. Dans une première étape, nous insérons les motifs liants que nous définissons comme *aromatiques*. Ce sont des cycles aromatiques complets ou partiels (arcs) qui sont insérés dans l'enveloppe en face des cycles aromatiques du substrat. En effet lorsque deux cycles aromatiques sont positionnés l'un en face de l'autre, leurs nuages électroniques se confondent et créent des liaisons.

La deuxième étape est l'insertion de motifs hydrogènes. Il existe deux types de motifs hydrogènes : les donneurs et les accepteurs. Ces deux types de motifs sont complémentaires, pour qu'une liaison hydrogène se crée un donneur et un accepteur sont nécessaires. Ainsi nous commençons par chercher les donneurs et les accepteurs présents dans le substrat afin de positionner un module complémentaire dans l'enveloppe. Un motif donneur est composé d'un atome hydrogène relié un hétéro-atome (atome qui n'est ni de l'hydrogène ni du carbone). La liaison hydrogène se fait à partir l'atome d'hydrogène. Un motif accepteur est un atome de carbone, d'azote ou de fluor qui possède au moins un doublet non-liant, bien que les atomes de fluor soient moins courants dans les liaisons hydrogènes. La liaison hydrogène se fait dans la direction d'un doublet non-liant.

Cependant il est difficile pour plusieurs atomes d'un même motif de participer à des liaisons hydrogènes en même temps. C'est pourquoi pour chaque motif hydrogène du substrat un seul motif hydrogène est inséré dans l'enveloppe. Si le motif hydrogène du substrat peut créer des liaisons hydrogènes à plusieurs positions différentes, alors plusieurs solutions seront générées, chacune d'entre elles utilisant une position différente. C'est par exemple le cas d'un donneur qui posséderait plusieurs doublets non-liants dans le substrat.

En conclusion, nous avons réussi à développer une approche et une application permettant la conception de guide de construction de cages moléculaires spécifique à un substrat donné.

## Résultats actuels

Les résultats actuels présentent plusieurs intérêts. Le fait de passer par une structure intermédiaire permet d'avoir un guide pour générer les cages moléculaires les plus spécifiques possible au substrat. En effet l'enveloppe a une forme complémentaire à celle du substrat, grâce à elle les informations géométriques du substrat sont conservées dans les cages générées pour une plus grande spécificité des cages. Elle permet de pouvoir capturer toutes les irrégularités de la forme du substrat qui ne sont pas forcément visibles à l'œil. Elle permet également de pouvoir définir les zones inaccessibles, c'est-à-dire des zones dans lesquelles on aurait aimé pouvoir placer des motifs moléculaires dans les cages mais qui s'avèrent être inaccessibles en tenant compte des

aspects chimique et géométrique du substrat.

La deuxième phase permet, quant à elle, de pouvoir positionner les motifs des cages moléculaires les plus importants puisqu'il s'agit des motifs liants qui serviront à créer des interactions entre le substrat et les cages générées. L'intérêt des motifs aromatiques est qu'ils épousent parfaitement la forme du substrat, on conserve ainsi une meilleure spécificité. Quant aux liaisons hydrogènes, ce sont les plus importantes des liaisons faibles. Les motifs hydrogènes permettent donc une meilleure capture du substrat par les cages.

## Perspectives

L'une des premières étapes de notre approche, que nous avons développé dans le chapitre 2, est la recherche de la topologie des atomes du substrat en fonction de leur géométrie dans l'espace. Nous avons défini un ensemble de règles permettant cette recherche. Cependant, en pratique les électrons des atomes sont libres. Il arrive ainsi que certains électrons passent d'une liaison de l'atome à une autre générant ainsi des altérations dans la topologie des atomes adjacents. Dans notre approche, si les deux sommets adjacents n'ont que le sommet auquel appartient cet électron comme voisin, nous leur trouverons la même topologie. En pratique, ils auront deux géométries différentes qu'ils échangeront au cours du temps. Une des perspectives est d'affiner le modèle afin de prendre en compte ces rares cas de figure. De plus, l'échange d'électron amène à des configurations différentes, plusieurs enveloppes pourraient donc être générées pour un même substrat en fonction de ces configurations.

Dans le chapitre 3, nous avons intégré les motifs aromatiques et les motifs hydrogènes. Une autre perspective est de rajouter l'intégration d'autres modules liants. Il existe en effet d'autres types de liaisons faibles telles que les liaisons ioniques qui sont des interactions électrostatiques entre ions ([Wikipédia \(b\)](#)), les liaisons halogènes ([Wikipédia \(c\)](#)) ou encore les liaisons de Van der Waals ([Wikipédia \(d\)](#)). Bien qu'elles soient moins courantes, certaines d'entre elles, comme les liaisons ioniques, ne sont pas rares. Rajouter d'autres types de modules liants pourrait être particulièrement intéressant dans les cas de substrats qui possèdent peu de liaisons aromatiques et hydrogènes afin de renforcer malgré tout les interactions entre le substrat et les cages générées.

Une autre perspective est de générer des graphes moléculaires complets comme guide de construction. En effet, notre approche permet la conception de guides comprenant les informations sur la forme générale que doivent suivre les cages moléculaires pour une meilleure spécificité, ainsi que les modules de liaisons nécessaires pour les interactions. La suite logique de cette approche est de poursuivre les intégrations de motifs moléculaires afin de lier les motifs déjà présents entre eux dans le but de construire des graphes moléculaires complets représentant les cages. Cet assemblage est très complexe et est un problème de recherche en lui-même. En effet, puisque plusieurs modules sont déjà positionnés, il faut tenir compte des écarts et des angles qui les séparent afin de respecter les contraintes chimiques des molécules. Les angles séparant les voisins d'un atome sont limités en fonction de sa géométrie, de même les écarts qui séparent deux atomes sont bornés. Pour relier deux motifs liants entre eux il faut donc trouver la bonne combinaison d'assemblage de motifs pour que les angles et les distances entre les atomes soient cohérents. La difficulté est d'autant plus importante

qu'on se trouve dans un espace 3D. De plus, il faut aussi prendre en compte la *zone interdite* qu'est l'enveloppe du substrat puisque aucun sommet d'une cage ne doit être positionné à l'intérieur. Et il faut également rester au plus proche de l'enveloppe pour conserver au maximum la spécificité des solutions générées.

Ensuite la rigidité de la solution finale est aussi à prendre en considération. Lorsque les atomes d'une molécule sont reliés par des liaisons simples, la forme générale de la molécule évolue constamment car les atomes peuvent pivoter autour de l'axe de la liaison. Plus une molécule est constituée de liaisons doubles ou triples et de cycles, plus elle est rigide et conservera sa forme. Il est donc important de choisir les motifs à intégrer de telle sorte que les cages qui pourront être construites à partir des graphes moléculaires générés soient le plus rigide possible.

En prenant toutes ces considérations en compte, l'idée serait de paver les espaces entre les différents motifs déjà présents en utilisant d'autres motifs moléculaires. Ce pavage doit être réalisé de façon à conserver l'emplacement des motifs liants autour du substrat tout en garantissant les angles et les distances entre les sommets ainsi que la rigidité de la solution finale pour que cette dernière soit réalisable d'un point de vue moléculaire.

# Chapitre 5

## Bibliographie

- Ahmad, N., Younus, H.-A., Chughtaiabd, A.-H., and Verpoort, F. (2015). Metal-organic molecular cages : applications of biochemical implications. *Chem. Soc. Rev.*, 44(9).
- Alexander, C., Andersson, H.-S., Andersson, L.-I., Ansell, R.-J., Kirsh, N., Nicholls, I.-A., O'Mahony, J., and Whitcombe, M.-J. (2006). Molecular imprinting science and technology : a survey of the literature for the years up to including 2003. *Wiler InterScience*, 19 :106–180.
- Alexiadou, D. K., Maragou, N. C., Thomaidis, N. S., Theodoridis, G. A., and Koupparis, M. A. (2008). Molecularly imprinted polymers for bisphenol a for hplc and spe from water and milk. *Journal of Separation Science*, 31 :2272–2282.
- Autodesk. <https://www.autodesk.com/>.
- Balaban, A. (1985). Applications of graph theory in chemistry. *J. Chem.Inf. Comput. Sci.*, 25(3) :334–343.
- Barth, D., Boudaoud, B., Couty, F., David, O., Quessette, F., and Vial, S. (2012). Map generation for co2 cages. *ISCIS 2012*.
- Barth, D., David, O., Quessette, F., Reinhard, V., Strozecki, Y., and Vial, S. (2015). Efficient generation of stable planar cages for chemistry. *14th International Symposium on Experimental Algorithms SEA 2015*, 9125 :235–246.
- Bonichon, N., Gavoille, C., Hanusse, N., and Ilcinkas, D. (2010). Connections between theta-graphs, delaunay triangulations, and orthogonal surfaces. *36th International Workshop WG 2010 : Graph-Theoretic Concepts in Computer Science*, 6410 :266–280.
- Brinkmann, G., Friedrichs, O. D., Liskien, S., Peeters, A., and Cleemput, N. V. (2010). Cage - a virtual environment for studying some special classes of plane graphs. *Match-Commun. Math. Comput. Chem. - an Update*, 63(3) :533–552.
- Brown, C. J. and Farthing, A. C. (1949). *Nature*, 164 :915–916.
- Byrne, K., Zubair, M., Zhu, N., Zhou, X.-P., Fox, D.-S., Zhang, H., Twamley, B., Lennox, M.-J., Düren, T., and Schmitt, W. (2017). Ultra-large supramolecular coordination cages composed of endohedral archimedean and platonic bodies. *Nature Communications*.

- Centre, T. C. C. D. <http://www.ccdc.cam.ac.uk/>.
- Chen, L., Wang, X., Lu, W., Wu, X., and Li, J. (2016). Molecular imprinting : perspectives and applications. *Journal of Chemical Society Reviews*, 45 :2137–2211.
- Cram, D. J. (1987). The design of molecular hosts, guests, and their complexes. *Nobel Lecture*.
- Cram, D. J. and Steinberg, H. (1951). *Journal of the American Chemical Society*, 73 :5691–5704.
- CRAN. Alphashape3d website. <https://CRAN.R-project.org/package=alphashape3d>.
- Delaunay, M. (1924). Sur la sphère vide. *Congrès international des mathématiciens*, pages 695–700.
- Dr, Z.-S. L., Zheng, C., Yan, C., and Gao, R.-Y. (2007). Molecularly imprinted polymers as a tool for separation in cec. *Electrophoresis*, 28 :127–136.
- Edelsbrunner, H. and E.P.Mücke (1994). Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13(1) :43–72.
- Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4) :551–559.
- Ellis-Davies and R, G. C. (2007). Caged compounds : Photorelease technology for control of cellular chemistry and physiology. *Nature methods* 4.8, page 619–628.
- Evans, W., Gansner, E., Kaufman, M., Liotta, G., Meijer, H., and Spilner, A. (2011). Approximate proximity drawing. *In proc. of Graph Drawing : 19th International Symposium, GD*, 46(6) :166–178.
- Friedrichs, O. D. (2000). Fast embeddings for planar molecular graphs. *Discrete Mathematical Chemistry, Am. Math. Soc.*, pages 85–95.
- Geldenhuis, W.-J., Malan, S.-F., Bloomquist, J.-R., Marchand, A.-P., and der Schyf, C.-J. V. (2005). Pharmacology and structure-activity relationships of bioactive polycyclic cage compounds : a focus on pentacycloundecane derivatives. *Med Res Rev*, 25(1) :21–48.
- Gillespie, R. and Nyholm, R. S. (1957). Inorganic stereochemistry. *Quart. Rev. Chem. Soc.*, 11 :339–380.
- Gillespie, R. J. (1963). The valence-shell electron-pair repulsion (vsepr) theory of directed valency. *Journal of Chemical Education*, 40(6) :295–301.
- Gokel, G. W. and Cram, D. J. (1973). *Journal of the American Chemical Society*, 92 :481–482.
- Hof, F., Craig, S. L., Nuckolls, C., Rebek, J., and Prof, J. (2002). Molecular encapsulation. *Angew. Chem. Int.*, 41(9) :1488–1508.
- Kutz, O., Hastings, J., and Mossakowski, T. (2012). Modelling highly symmetrical molecules : Linking ontologies and graphs. *LNAI proc. of 15th International Confe-*

- 
- rence on *Artificial Intelligence AIMSA-12*, 7557 :103–111.
- Lehn, J.-M. (1987). Supramolecular chemistry - scope and perspectives molecules - supramolecules - molecular devices. *Nobel Lecture*.
- Lehn, J.-M. (1999). Dynamic combinatorial chemistry and virtual combinatorial libraries. *Chemistry, A European Journal*, 5(9) :2455–2463.
- Liu, C.-Y. and Lin, C.-C. (2004). An insight into molecularly imprinted polymers for capillary electrochromatography. *Electrophoresis*, 25 :3997–4007.
- Mastalerz, M. (2010). Shape-persistent organic cage compounds by dynamic covalent bond formation. *Angew. Chem.*, 49(30) :5042–5053.
- Mastalerz, M. (2012). Permanent porous materials from discrete organic molecules-towards ultra-high surface areas. *Chem. Eur. J.*, 18(33) :10082–10091.
- Mitra, T., Jelfs, K., Schmidtman, M., Ahmed, A., Chong, S. Y., Adams, D. J., and Cooper, A. (2013). Molecular shape sorting using molecular organic cages. *Nature Chemistry*, 5(4) :276–281.
- Mosbach, K. (1994). Molecular imprinting. *Trends in Biochemical Sciences*, 19 :9–14.
- Pedersen, C. J. (1967). *Journal of the American Chemical Society*, 89 :7017–7036.
- Pedersen, C. J. (1987). The discovery of crown ethers. *Nobel Lecture*.
- research group at Iowa State University, G. <http://www.msg.chem.iastate.edu/gamess/>.
- Schrödinger. <https://www.pymol.org/>.
- Sidgwick, N. and Powell, H. (1940). Bakerian lecture. stereochemical types and valency groups. *Proc. Roy. Soc.*, 176(965) :153–180.
- Song, X., Xu, S., Chen, L., Wei, Y., and Xiong, H. (2014). Recent advances in molecularly imprinted polymers in food analysis. *Applied Polymer science*, 131.
- Tsuchida, R. (1939). New simple valency theory. *Journal Chem. Soc. Jpn*, 60(3) :245–256.
- Whitcombe, M. J., Kirsh, N., and Nicholls, I. A. (2014). Molecular imprinting science and technology : a survey of the literature for the years 2004–2011. *Journal of Molecular Recognition*, 27 :297–401.
- Wikipédia. [https://fr.wikipedia.org/wiki/Format\\_.xyz](https://fr.wikipedia.org/wiki/Format_.xyz).
- Wikipédia. [https://fr.wikipedia.org/wiki/Liaison\\_ionique](https://fr.wikipedia.org/wiki/Liaison_ionique).
- Wikipédia. [https://fr.wikipedia.org/wiki/Liaison\\_halogène](https://fr.wikipedia.org/wiki/Liaison_halogène).
- Wikipédia. [https://fr.wikipedia.org/wiki/Force\\_de\\_van\\_der\\_Waals](https://fr.wikipedia.org/wiki/Force_de_van_der_Waals).
- wwPDB Foundation. <http://www.wwpdb.org/index>.
- Zhang, G. and Mastalerz, M. (2014). Organic cage compounds - from shape-persistency

to function. *Chem. Soc. Rev. A*, 43(6) :1934–1947.

# Annexe A

## Rappel de chimie : Règle du duet et règle de l'octet

Il existe une famille d'éléments chimiques appelés *gaz rares* ou *gaz nobles*. Les éléments de cette famille n'existent qu'à l'état de gaz monoatomique car ils ont la particularité d'être extrêmement stables, c'est-à-dire que leur couche externe est totalement remplie (saturée).

Les autres atomes ont un défaut ou un excès d'électrons par rapport aux gaz rares, ce qui leur confère une structure électronique instable. Ils cherchent à acquérir une stabilité grâce des transformations. Chaque transformation tend à les rapprocher de la structure électronique du gaz rare le plus proche dans la classification du tableau périodique.

Un ion est un atome qui a subi une ou plusieurs transformations. Un ion est stable si sa couche électronique externe est saturée :  $(K)^2$  ou  $(L)^8$  ou  $(M)^8$

Les règles du duet et de l'octet s'appliquent aux atomes des éléments chimiques des trois premières lignes du tableau périodique. Elles permettent de prévoir quelles transformations sont nécessaires à un ion donné pour avoir une structure électronique stable.

Deux cas sont à distinguer : le cas où le numéro atomique de l'atome est inférieur ou égal à 4, et celui où il est supérieur à 4.

### Règle du duet

La règle du Duet s'applique aux atomes de numéro atomique  $Z \leq 4$ , en l'occurrence l'hydrogène, l'hélium et le lithium.

**Definition 24.** *Un atome ou un ion est stable si sa couche externe est la couche  $K$  et est remplie avec deux électrons.*

Le gaz rare de référence est l'hélium qui a un numéro électronique  $Z = 2$ . L'hydrogène dont  $Z = 1$  cherche donc à saturer sa couche ( $K$ ) en acquérant un électron, formant ainsi l'ion  $H^-$ . À l'inverse le lithium ( $Z = 3$ ) cède un électron pour former l'ion  $Li^+$ .

---

## Règle de l'octet

La règle de l'octet s'applique aux atomes de numéro atomique  $Z > 4$ .

**Definition 25.** *Un atome ou un ion est stable si sa couche externe ( $L$  ou  $M$ ) est remplie avec huit électrons.*

Deux gaz rares peuvent servir de référence : le Néon ( $Z = 10$ ) et l'Argon ( $Z = 18$ ). Les autres éléments chercheront à se rapprocher de leur structure en acquérant ou cédant des électrons.

Par exemple, un atome de carbone ( $Z = 6$ ) cherchera à acquérir quatre électrons en formant quatre liaisons avec d'autres atomes : quatre liaisons simples, ou une liaison simple et une liaison triple, ou deux liaisons doubles, ou encore deux liaisons simples et une liaison double.

## Annexe B

# Étude de l'influence du paramètre $\alpha$ de la méthode alphashape

Comme expliqué dans le chapitre 2, la méthode de l' $\alpha$ -shape dépend du paramètre  $\alpha$ . Ce paramètre représente la taille de la sphère de référence. Les rayons des sphères circonscrites aux tétraèdres de la triangulation 3D sont comparés à  $\alpha$ . Si le rayon de la sphère est supérieur à  $\alpha$ , les faces du tétraèdre sont retirées de la triangulation.

Si le paramètre  $\alpha$  est trop petit, nous n'obtenons pas une enveloppe. L'exemple de la Figure B.1 montre le résultat de l'enveloppe obtenue avec  $\alpha = 1$  sur la molécule d'adénosine. Ici, on peut voir que nous n'obtenons ni une enveloppe, ni même des triangles.

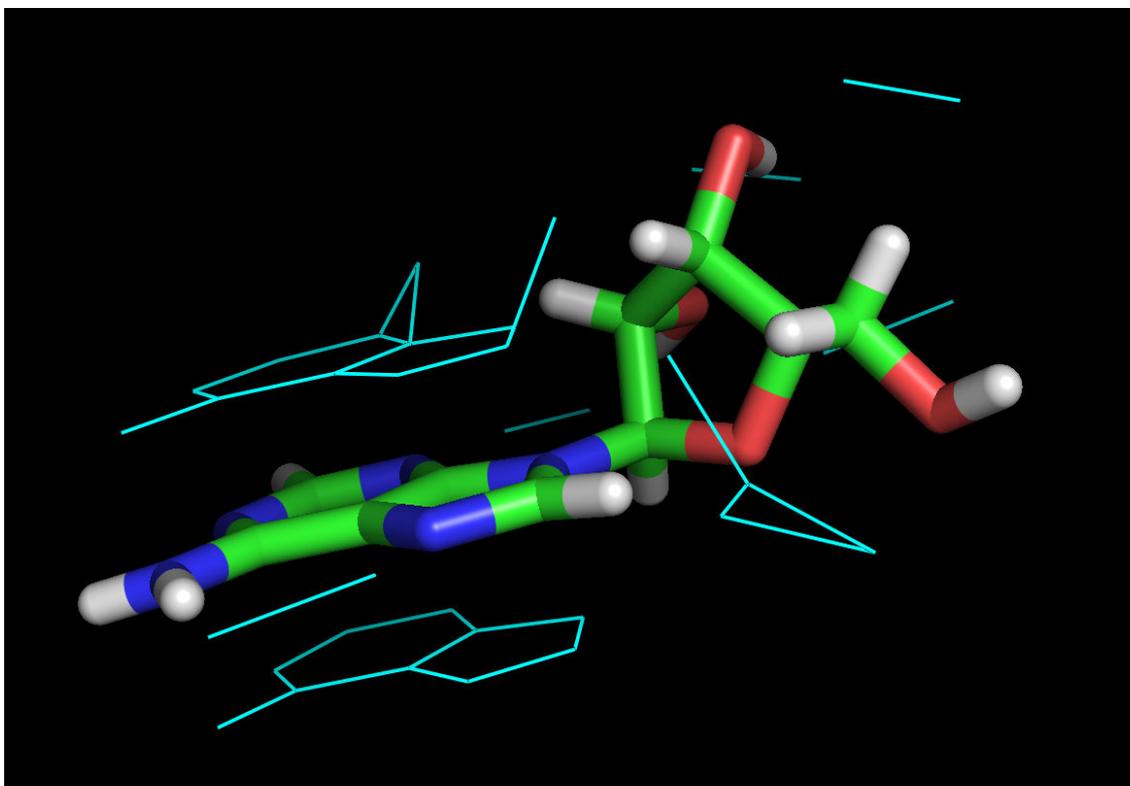


FIGURE B.1 – Enveloppe de la molécule d'adénosine obtenue avec  $\alpha = 1$ .

Sur la Figure B.2, nous pouvons voir le résultat de l'enveloppe, toujours de la

molécule d'adénosine, avec  $\alpha = 2$ . Du fait que la sphère de référence est trop petite, certains triangles qui devraient former l'extérieur de l'enveloppe sont supprimés. Ainsi plusieurs triangles intérieurs sont référencés comme étant à l'extérieur. En conséquence, certaines arêtes de l'enveloppe traversent le substrat.

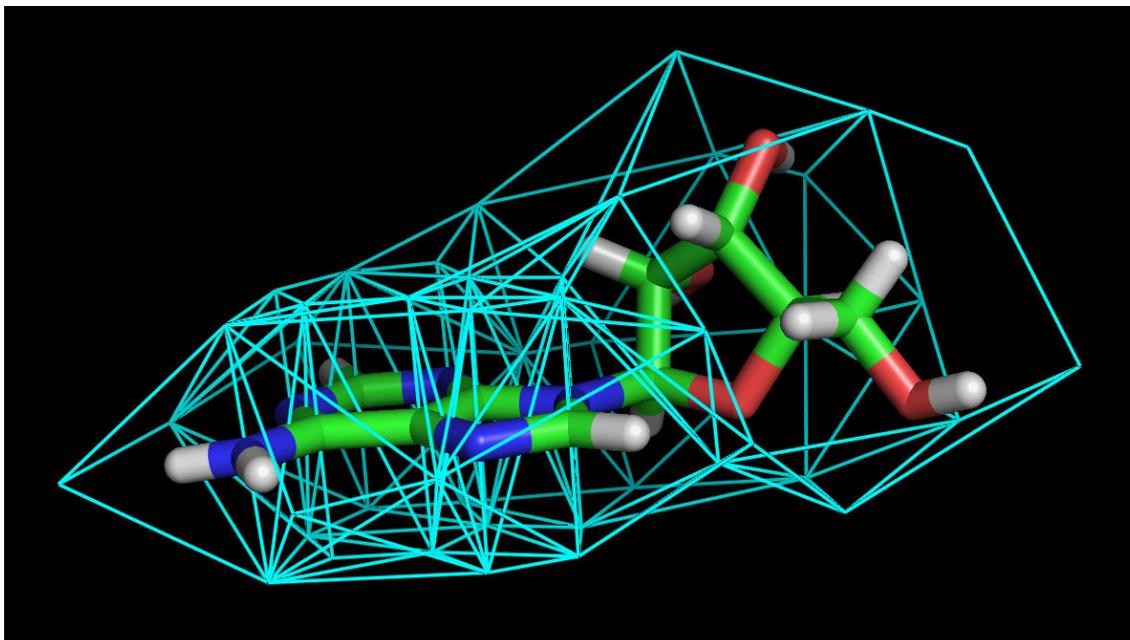


FIGURE B.2 – Enveloppe de la molécule d'adénosine obtenue avec  $\alpha = 2$ .

En testant sur une quarantaine d'exemples, nous avons pu observer que à partir d'un  $\alpha$  égal à 3, nous obtenons des résultats exploitables quel que soit le substrat. Nous avons donc défini la borne minimum du paramètre  $\alpha$  à 3. Il n'y a pas de borne maximum. Si l'alpha est trop grand, l'enveloppe obtenue sera l'enveloppe convexe.



**Titre :** Modélisation et Algorithmique de graphes pour la construction de structures moléculaires

**Mots clés :** Algorithmique, molécules, graphes, modélisation

**Résumé :** Dans cette thèse, nous présentons une approche algorithmique permettant la génération de guides de construction de cages moléculaires organiques. Il s'agit d'architectures 3D semi-moléculaires possédant un espace interne défini capable de piéger une molécule cible appelée substrat. De nombreuses œuvres proposent de générer des cages organiques moléculaires obtenues à partir de structures symétriques, qui ont une bonne complexité, mais elles ne sont pas spécifiques car elles ne prennent pas en compte des cibles précises. L'approche proposée permet de générer des guides de construction de cages moléculaires organiques spécifiques à un substrat donné.

Afin de garantir la spécificité de la cage moléculaire pour le substrat cible, une structure intermédiaire, qui est une expansion de l'enveloppe du substrat cible, est utilisée. Cette structure définit la forme de l'espace dans lequel est piégé le substrat.

Bien que cette structure ne soit pas assimilable à une molécule, elle est construite dans une logique

moléculaire. Chaque sommet qui la compose est situé dans une direction dans laquelle l'un des atomes du substrat peut créer des interactions avec des atomes d'autres molécules. Pour ce faire, nous utilisons les propriétés des éléments chimiques qui composent le substrat, ainsi que la méthode de VSEPR. Cette méthode permet de prédire la géométrie des atomes en se basant sur la théorie de la répulsion des électrons. À partir de leur géométrie, pour chaque atome du substrat, on peut retrouver les directions des doublets non-liants d'un atome. Les directions de ces doublets non-liants sont les directions dans lesquelles un atome peut créer des interactions (liaisons) avec d'autres atomes.

Des petits ensembles d'atomes, appelés motifs moléculaires liants, sont ensuite intégrés à cette structure intermédiaire. Ces motifs moléculaires sont les ensembles d'atomes nécessaires aux cages moléculaires pour leur permettre d'interagir avec le substrat afin de le capturer.

**Title :** Modelling and graph algorithms for building molecular structures.

**Keywords :** Algorithmic, Molecules, Graphs, Modelling

**Abstract :** In this thesis, we present an algorithmic approach allowing the generation of construction guides of organic molecular cages. These semi-molecular architectures have a defined internal space capable of trapping a target molecule called substrate. Many works propose to generate molecular organic cages obtained from symmetrical structures, which have a good complexity, but they are not specific because they do not take into account precise targets. The proposed approach makes it possible to generate guides for the construction of organic molecular cages specific to a given substrate.

In order to ensure the specificity of the molecular cage for the target substrate, an intermediate structure, which is an expansion of the envelope of the target substrate, is used. This structure defines the shape of the space in which the substrate is trapped. Although, this structure is not comparable to a mole-

cule, but it is built in a molecular logic. Each vertex which composes it, is located in a direction in which one of the atoms of the substrate can create interactions with atoms of other molecules. To do this, we use the properties of the chemical elements that make up the substrate, as well as the VSEPR method. This method predicts the geometry of atoms based on the electron repulsion theory. From their geometry, for each atom of the substrate, we can find the directions of the lone pairs of an atom. The directions of these lone pairs are the directions in which an atom can create interactions (bonds) with other atoms.

Small sets of atoms, called molecular binding patterns, are then integrated into this intermediate structure. These molecular patterns are the sets of atoms needed by molecular cages to allow them to interact with the substrate to capture it.

